UNIVERSITAT ROVIRA I VIRGILI

Universitat Oberta de Catalunya

Master in Computational and Mathematical Engineering

Final Master Project (FMP)

Machine learning techniques for computational Marketing into Energy Sector

Author: Miquel Rius Carmona

Supervisors: Dr. Eva Vallada and Dr. Angel A. Juan

15/06/2018

Signature of the director authorizing the final delivery of the FMP:

Index

1. Introduction

1.1 Context and justification of the Work

1.1.1 Introduction

In the last years, the energy sector in Spain has been involved in a growing movement in the number of enterprises that provides commercial services.

About twenty years ago the energy sector was limited due the own energy distributors company were who also sell the energy to the final consumers. This competence limitation generates a soft market where innovation was not necessary, and every enterprise has his own piece of share in the sector.

Later, Spain government opened the wall, and the legal difficulties to create an energy company were inexistent, so a high number of marketers appears.

Until a few years, despite the increased competence the sector was very classical in what and how the companies sold to the customers.
This companies were limited in trade energy basing in OMIP prices (nowadays and future). This means that a marketer has a final goal, buy to the cheapest possible price and sell it higher to achieve some profits.

As every sector, the technological evolve affects this company and what before were mechanical energy counters now we are talking in smart-counters. The typical house nowadays is a smart house. The transport is suffering a big change and not only are using fossil fuel, but also gas and electricity.

The apparition of the Internet of Things, the smart cities, social networks, artificial intelligence and others create a different environment about who are the final customers, and what they need.

These changes generates a highly flexible company able to understand quickly what customers want and need.

1.1.2 Computational Marketing

The field that assist with marketing strategies is an emerging field called Computational Marketing. "Computation is a way to give a form and function to blocks of information too large for any person to analyse" (Marketing-Schools, 2012).

What generates this data are different sources, for example, social networks, devices, websites and the user itself.
The goal of computational marketing (CM) is not to create, but to connect customer with products, advertises and any other marketing field.

This area has a lot of advantages, but "the greatest advantage of computational marketing is that it automates many of the traditional duties of marketing". This means that a lot of functions developed by a marketing agent now can be automated, allowing this agent to invest their time to more important, and high value tasks.

CM draws on several disciplines within computer science, and economic theory. The main components are:

- Information retrieval: To analyse data, first of all you need this data.
- Machine learning: Artificial intelligence branch that creates ways to recognize and react to complex patterns.
- Optimization: On one hand, is important to maximize the accessibility of a user into a website, but also machine learning uses **heuristics** that involves optimization of a goal function.
- Microeconomics: Science that studies economy in a small scale. For example, study the way individuals make choices about what to buy.

It is true that enterprises with a strong online presence will have better benefits of computational marketing, but everyday there is more public information about customers that any enterprise can use and manipulate to extract patterns and improve their business policies in marketing.

A great deal is that CM has only been possible for few years, because before the servers does not have enough power to run some algorithms. This means, that there is an enormous potential for the field.

1.1.3 - Machine learning

Here is where machine learning appears.

The technological changes not only affect companies in what they offer, but also in their functionalities. Artificial intelligence and more specifically Machine Learning (ML) open a window of opportunities to become more flexible and quickly identify what a customer need and offer it to them. In this field, the developers teach machines to do things. There are mainly three possible ways of learning through data:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

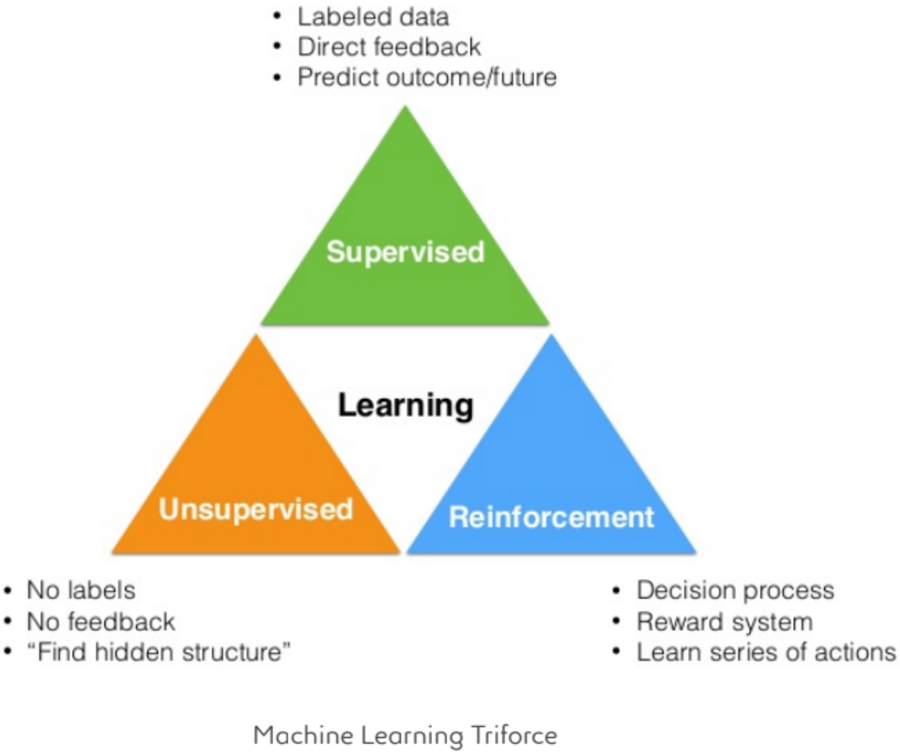


Figure 1: Machine Learning Triforce learning methods

1.1.3.1 Supervised Learning

It is the most common field. It consists in learning a decision model from data thanks to labels associated with data. A big number of Data Science models uses this technique to develop their models.

As is told in the Machine Learning for Grandmas article a good example is: "We have a dataset of many images of cats and dogs. For each image we know if it represents a cat or a dog, and these images and their labels are given to the model. It should be able to make the difference on its own. This means that if all goes well, when showing a new image without label, the model can tell us if it is a dog or a cat."

In the image under this line you can find a schematic summary of the methodology.

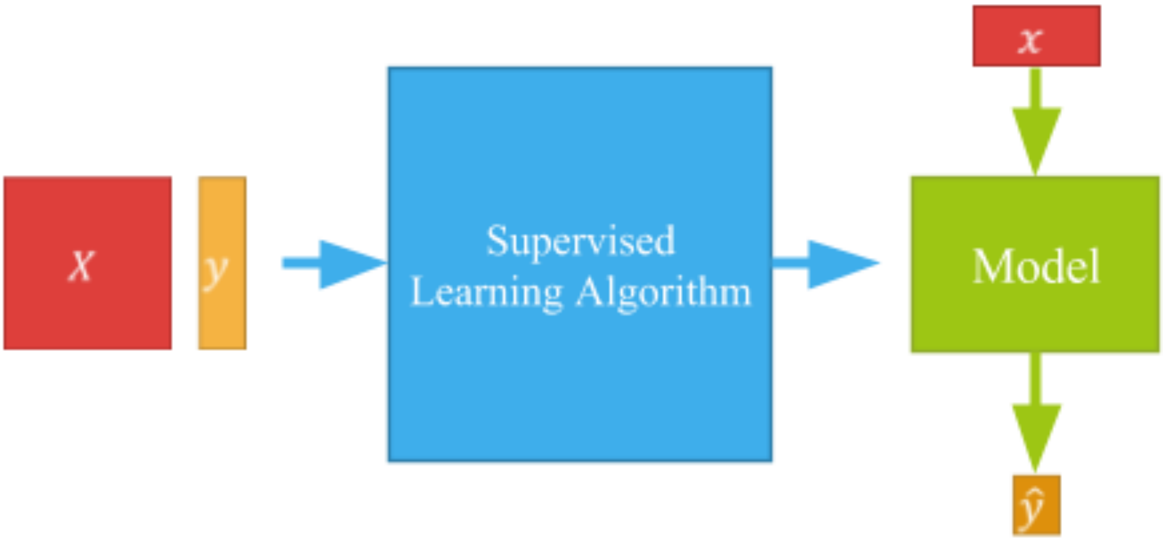


Figure 2: Supervised Learning Algorithm Schema

The red squares are the images send to the model. The yellow ones are the labels. The Supervised Learning Algorithm (SLA) tries to learn how different X characteristics effects on Y label. When it is trained with enough data, it will be able to tell you the label (y) for the X that you would like to classify. This is possible because the model contains a set of rules learned through training dataset, then it applies these rules to any input, and gives you the computed output.

1.1.3.2 Unsupervised Learning

Unsupervised Learning is the second big field in ML. This has a main difference between the previous ML technique. In this case it does not have labels. So, the algorithm will work blind to find a structure in data that it could interpret.

This methodology will not be two ways process like in SL. The algorithm does not learn on some data to predict the label for future data. In this case, the algorithm creates the structure finding feature values of the different input data.

Still using Machine Learning for Grandmas article examples: "Consider same dataset of dogs and cats but without labels. So, the machine does not know what they represent.

The algorithm will be able to find the underlying structure of the dataset of images defines two groups of similar images (clusters). Each data point (image) would be assigned to a cluster. Finally, using specific domain expertise of cats and dogs, we could define that one cluster contains cats images and the other contains dogs images."

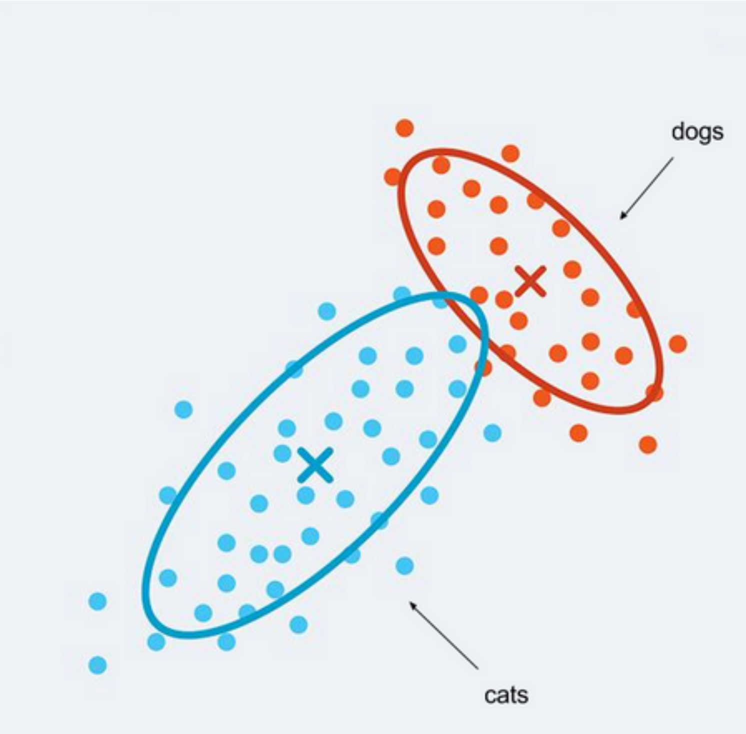


Figure 3: Clustering images

It is very important to remark the affirmation of "specific domain expertise". This means that for this methodology we do not have any ground truth. So, it could be a little bit confusing, and the better way to take profit is to use the human expertise in technical field. This get big importance to understand that technologies do not substitute humans but facilitates some hard tasks to let humans focus on more profit tasks, for example, decision making.

To check these clusters then we need some expert, and not just more input data. This is because we do not know the label that we should obtain.

ULA could be used combined with SLA. For example, the first are used as a kind of data pre-processing step to have a better representation of data. Then, it is sent to a Supervised Learning algorithm.

Despite this, are also use cases where ULA can be used alone:

- Outlier detection: Banking sector, use these algorithms to mark as a risky customer the ones that does not enter in any cluster. This helps to prevent fraud.

- Customer segmentation: This means that we get the possibility to treat different our customers taking account of how they are. This is possible through the fact that the analysed data tell us the customer's behaviour. This helps to customize the offer to each customer.

This last point is very important for the project in fact. As seen later on, the project will have to create a customer segmentation to classify the customers of a company to later offer a better product for each customer.
This means **clustering, a technique that is used in Unsupervised Learning Algorithms, and will be the core of the project.**

1.1.3.3 Reinforcement Learning

Finally, there is another branch of Machine Learning. Despite Supervised and Unsupervised Learning cover the big part of the ML algorithms the Reinforcement Learning (RL) techniques are getting some fame due to the fact that their main functionality is in robot training.

This technique is base in to reward the good decisions, and to punish the wrongs. For every wrong decision, the robot learn that it is not the way, so it tries another one. Finally, when it gets a reward it knows a way to do this.

Autonomous cars learn using these techniques. Every time that they take a decision and crash (punish) they know that it is not the correct way.

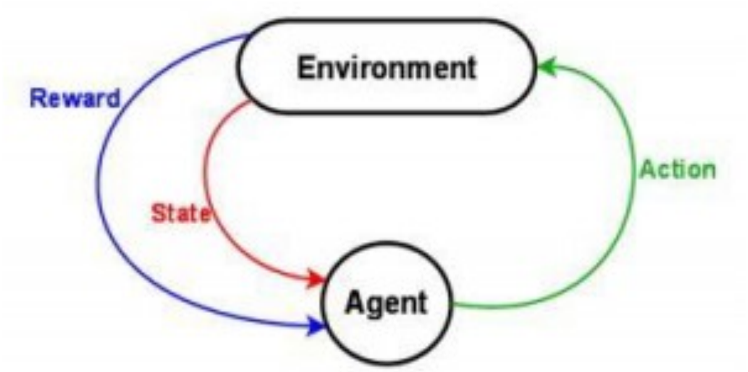


Figure 4: Reinforcement Learning Schema

Despite this branch is normally associated to robots as is told before, it could be use in any other sector or field.
For example, online advertising, or product generation. This is possible because, RL is like an AB-testing. Always reject the loser option and continues with the winner one.

The difficulty of this branch lies in the fact that actions taken by the agent might change the environment, and the agent might get its reward a long time after its action. In these cases, the agent loses the traceability of which reward is associated with an action itself.

1.1.3.4 How is the product generation into the energy sector

The product generation process studied will focus when a customer asks for an offer to an energy marketer enterprise.

The company tries to gather all the feasible data of the customers (using outside and inside sources) and study the case in question. To do this job, the company need a lot of resources.

1. Data scientists/engineers who retrieve the data and gather it.
2. Call centre or commercial agents who speaks with the customers.
3. Industrial or energy engineers who analyse the case and create the product.

All of them must have a great communication between them and the whole process must be fast enough to ensure that the customer does not goes to any other company.

This can be extrapolated to a clustering techniques. This means that to solve the first part of the problem is possible to apply a front-end software who asks some key questions to the customer (with a previous analysis of which questions should the company be asking to ensure the correct customer classification).

1.1.3.5 How can ML help into the energy sector?

As is possible to see, is highly important then to be one of the first applicants into machine learning algorithms to take advantage of it and be able to change quickly as market does.

This is notable because the energy companies have big departments of technical engineers who study the needs of a customer and creates the more accurate offer for them. Also respecting some business needs and restrictions.

Furthermore, is important to add the marketing and call centre department who individualize the treatment of each customer. This generates a high monetary cost in salary, but also in time and even in opportunity cost, because while these engineers must study every variable of a customer are not doing something more interesting for the company.

Machine learning techniques can help companies to create a more customized study of customer and grouping them in characteristics that for engineers are a high number of times forgotten.
This will reduce study time, manual work, and a better data normalization for the companies who applies these algorithms.

Moreover, will add high knowledge of who they customers are and what they want. And will facilitate product cross selling as insurance, maintenance and others from the main sold product. This will help to focus efforts in those customers with higher probabilities to acquire any other product of the company basing the prediction in how they are, and what has purchased other customers like them.

1.1.3.6 Methodological contributions of the Project

This project creates a new complete methodology of product generation in the energy sector using an already existent clustering library (Python), and an optimization algorithm approach used to solve others problems. The innovation of this project is to combine both independent works in one methodology applied to an specific sector which not yet is using this technology stack.

The real aim is to create a modular project that combine clustering algorithms with optimization algorithms. In this specific case the optimization algorithm is knapsack algorithm fully developed by the researchers that create this project. This is interesting because the knapsack algorithm is where the product generation is done, so in the future a complete self development help to add more restrictions and business rules. Otherwise, clustering is a very studied branch of data science where is hard to add value.

This work add value in the energy sector because there is no development yet related with data science applied into the computational Marketing branch. In fact, a lot of companies do not know yet that Computational Marketing exists and this creates a lack of investment in this are. Then, this project firstly will open the scope of what data science, and algorithms can do, not only in the customer analysis area. Additionally, the project has the hope to convince the companies to invest in developers that performs these developments. Do not forget that neither any project in the literature review contemplate the possibility to use optimisation algorithms to satisfy some heuristic.

This project then, innovates in this domain. It uses a very old optimisation approach, the knapsack problem solution to add the higher value into the product that will be offered to the customer.

1.1.3.7 Relationship between the project stack and master courses.

The computational and mathematics engineering master offers an interdisciplinary formation between applied science and engineering. The main goal of this master is to prepare the students for R+D positions.

As explained before the project develops a new methodology in one specific area, but to do that this has required a research job, both in the business sector and in the engineering area.

Once the research has been done a set of business restrictions, and different ideas applied into other projects from some researchers has been discovered, but to apply that ideas is important to choose the best technology stack.

The project is related with machine learning concepts because some techniques that the algorithm will use, an example is clustering. The algorithm will learn how to classify customers in different types. This clustering concepts are part of Artificial Intelligence, this is one of the subjects done in the master.

The best programming languages to apply these algorithms are Python and R. In a scientific scenario is more efficient and compute faster R. Otherwise Python is a little bit more inefficient but is a full stack technology. This means that with Python is feasible to develop a website. Specifically, Django framework has been used in this project.
Then, Python has been chosen thanks to the flexibility that it presents, and with the purpose to develop a complete and online platform that uses Artificial Intelligence Algorithm the only possible option to consider is Python. The Django framework is an easy election. It is the most used Python framework for website development, and it have an easy learning curve.

This development can be considered in the Data Science branch, so it is important to have data to analyse. Then, the next decision was to choose the best option to store and manage data. In another projects the researchers use some flat files ".txt", this helps with faster development because there is no need to set up a server, a database and neither is needed to learn the main database administration tasks.

Despite that, a database is a best option when there is a lot of data, or the structure will vary a lot, or it is not sure how data will increase. Moreover, the researcher in this project has working experience with database setup, administration and data management so the weakness of it use disappears, and only remains the advantages of easy data maintenance, and data manipulation. Then, the technology used is PostgreSQL, one open source database that has more computation power than any other open source database.

This tech stack has been studied in the master, Python as a programming language for data science, and PostgreSQL as an SQL database to manage data when the project not is about big data, in this case is recommended to use NO-SQL technologies.

1.2 Aims of the Work

1.2.1 General goals

Using as support the bibliography guide, and the techniques developed by other researchers and experts in similar fields, the purpose is to **develop a machine learning algorithm in Python to help energy sector enterprises in the classification of their customers and in the automatic product generation for these. These customers will be classified in different clusters using definition customer variables, to finally assign a customized product to satisfy customer needs.**

The algorithm will execute a knapsack optimization algorithm to find all the possible combinations of supervises that can fulfil a product with the restrictions specified by the customer.
As this combination will have a lot of feasible points, the output will be a CSV file with the cheapest product for the customer, the expensive one (understand that expensive cost means more profit for company) and finally the product with better ratio *"total price/Number of subservices"*.

1.2.2 Partial goals

To achieve the main objective of this project is interesting to define partial goals that will contribute in the correct achievement of the main purpose:

1. <u>Study the automatic generation product problem and the respective theorical and practical techniques developed by other researchers:</u>
   In the initial phase, is important to study previous investigation and business projects with the purpose to spot the strengths and the weakness of any other previous job done and also the key of his success. A correct study of the previous bibliography will help the researcher to does not commit the same mistakes and improve the final job.

2. <u>Process definition to problem solve:</u>
   The goal is to use the knowledge acquired in the previous phase and the experts interviewed inputs to create a correct method that propose a new or improved technique to solve the problem.

3. <u>Algorithm development:</u>
   This phase is the more important due the weight it represents in the main goal. It will consist in the development of an algorithm that using clustering and unsupervised learning techniques (machine learning) creates a customer classification.

   Later, this cluster data will be used to create the best fitting product for this cluster, but not only this product will be generated. Also, a product for the specified customer data will be created.
   This double product development allows to compare the differences that algorithm creates when using central point data (cluster) and customer data.

   To achieve that, the algorithm will use some business inputs and customer inputs. Also, will use some restrictions for product generation derivate for the real-life simulated environment where the algorithm would like to be tested. If these restrictions can't be covered, this means that the algorithm does not have a possible real-life application, at least how it has been interpreted and created.

   Then, it will create the best fitting product for the customer, and for the company. On the second case, we can understand the best product for the company as the expensive one for the customer. Higher cost should mean higher benefits for company (in real life it cannot apply).

On the other hand, the best fitting product of the customer will be or the one that considering all customer needs, have the cheapest price, or the one that considering all customers needs offers better ratio between total cost and the number of sub services offered.

4. Database development

   The objective of this phase consists in the correct creation of a database with the easiest data model possible that includes all the related data and variables for the correct execution of the algorithm. The key is to create the database small and flexible enough to be included in any real company, to ensure an easy and flexible deployment of the algorithm in production.

5. Design and create a set of experiments

   These experiments will help to check the methodology potential to solve the problem. Through the modification of some inputs, it will test the power and sensibility of the used techniques to solve real life problems. This will help to check if this is a good implementation and extracting a set of conclusions

1.3 Literature review and approach

Due to the lack of previous work in both fields: automatic product generation using machine learning and his application in energy sector. The approach used in the research will be innovative.

1.3.1 - Literature review

“Computational marketing is an emerging field that uses the power of computing to assist with marketing strategies”.

(http://www.marketing-schools.org/types-of-marketing/computational-marketing.html).

This simple sentence refers to whom computational marketing aims. Companies around the world in the future years will start to investigate and invest resources to this interesting field. Related to this investment companies will acquire:

1. Better ad relevance

2. Social Media and Behavioural

3. Reducing Costs

4. Automated Marketing Models

5. Harvesting User Information
6. Multi-Platform Analysis

Using this points that marketing-schools refers in his website, and the main trends in big data, business intelligence and data engineering is clear the motivation that moves this project forward.

The possibility to engage better business results with IT development using data is a huge motivation because it combines the best of business, technology and investigation world.

Despite that, computational marketing only exists a few years ago, and the lack of tech profiles, and technological resources on companies made that it cannot be exploded yet. Then, it is important to do an extensive literature review to check what other developers and investigators achieved.

Despite the innovation other researchers has developed previous work that can be a source of inspiration due it could be understood has a module of this project.

Some years later the concerns still the same but the methodologies have evolved.
Kengpol and Wangananon developed a neural network system to identify patterns in customers during the new product development projects.
Gu et al. applied also neural network to analyse the aesthetics evaluations given by users.
Han et al. considered to study the relationship between customer satisfaction and design variables using statistical regression.

All developers converge their opinions in the data entry. If data gathering is not correct, all the later on process will be unsuccessful even you use best technologies, resources and developers.

At 2014, E. Lutters, F. Van Houten, A.Bernard (Lutters, Van Houten, Bernard, 2014) and some other great investigators focus on tools and techniques for products design, and after a big investigation, the conclusion was that one of the tools is the data used in the design process.

One year later, at 2015, Zhenyu Zhang, Qingjin Peng, Peihua Gu (Zhang, Peng, Gu, 2015) develop an study related in how should a customer involved in product design.

Both studies arrive to the same conclusion you are creating products for customers, this data should be provided by them, but at least you design good questions, you will not achieve value from their answers. So, it is important to create good surveys, and speeches to achieve a great willingness from customers in product design process.

Furthermore, we should think that customers do not know that their opinions could be creating the new next star product in company.

This makes clear that, the data gathering process represents a big percentage of the project success.

Once you can achieve a good data gathering workflow, and using online tools such web technologies you could achieve that a company can receive constant feedback from products through online surveys for example, and then analyse data that will provide the pros and cons of a product, providing the company with data to create new products, or selling strategies (Lee,BradLow, 2011).

Within this we enter on data analysis, Salman Nazari-Shirkouhia,Abbas Keramati, just two years ago develop an algorithm to model customer satisfaction related to a new product. To do this, he used answers from customers to specific questions designed by marketing team that allows to understand the level of satisfaction to determined properties of products.

Is very important to understand that if you do not ask, for example, for colour opinion or you create ambiguous questions, the answers cannot provide feedback of this property, or will provide vague feedback of it. Then, we recover what was told previously, despite the algorithm uses the best possible techniques, will be impossible to model the satisfaction (in this study) of customers against this specific feature.

Then, is possible to see that all stages in this project will inherit the quality of the previous one, arriving initially to the data quality. All the studies focus their efforts to one of the stages, but has been possible to found one real company project, at 2009, that create a complete platform for product design in an engineering company, (Kuang, Jiang, 2009). Their purpose was “to create a platform and individual parameters are identified, and the quantified relationship between the product's perceptual image and the design parameters is established by using regression analysis from an affective evaluation survey.”

To achieve that customers were grouped by a cluster analysis of the. Based on the clusters, the number of platforms is determined. Then, using the similarity between individual preference and cluster preferences stablish a relationship. Finally, the values of the individual parameters are determined based on the satisfaction of each customer group”.

The key point of all the research mentioned above were:

1.Algorithmic research
2.Input data about what customers thinks.
3.Customer happiness
4.Feasible change improvement of products.

After analysing carefully what ideas are good of these researches for the here present, is possible to identify the following concepts.

For every measurable variable, the researchers have developed a set of customer analysis questions. These questions allow to understand how the happiness in this specific variable is. This means to involve the customer in the product design with a set of surveys.

Other researchers have studied a more procedural view in the product design. For example, Eric Lutters et.al have investigated a different set of tools and techniques for product design. They stand out that to study products is important to classify them. To facilitate this task is important to select a set of appropriate properties.

Choose those properties is not an easy job, because the determination of the quality of products in general is subjective.

After reading and check the whole literature, the project aim is to achieve a complete online based platform using web technologies that allows to gather constant user information through specific designed surveys that helps in product generation.

The editorial PACKT publishing has a lot of free eBooks (Richert,Coelho - 2013), (Romano, 2015), (Gollapudi, 2016), (Hearty,2016), (Cánepa, 2016) that has helped a lot with the project development of machine learning algorithm using Python functions and some other difficulties related with the development process.

1.3.2 The approach

1.3.2.1 Clustering data

The project will use machine learning as the algorithm central key. Specifically, Unsupervised Learning Algorithm (ULA).

The ULA approach will help companies in the Customer Classification Task (CCT). This algorithm will receive as an input the customer data:

- Tariff: Tariff of the customer. It is related directly with electric power and consumption.
- Customer type: Enterprise or home service.
- Hired power: The power hired for the customer. When you use more than this hired power, the marketer applies an extra cost. Is a powerful tool of cost optimization for the companies?
- Consumption: The real monthly consumption.
- Product type: In the electrical sector, mainly, are two possible products. One is at a fixed price, and another is following the real cost of market.

This data will classify the customer in clusters. In the results will be seen which variables presents more weight in customer clustering. It is easy to imagine which of these will provide the “definition” of a cluster.

1.3.2.2 Optimisation data

Furthermore, the research will develop a set of questions to know the interest of a customer for a specified sub product:

- Electrical car: It provides fast-charging points. Interesting for electrical cars and any other electric vehicle.
- Battery: It is related with the hired power. As told before when you use more than the hired power, the company will charge to your bill an extra cost. This can be solved if you include a battery that saves all the power that you do not use, and later, when you have a peak and get over you will use this energy warehouse and not the provided by the company. Then, they cannot charge you a plus.
- Smart home: Domotic houses have higher electric consumptions and must ensure that all the domestic stuff is properly installed. This can be done by a marketer company to improve the energetic efficiency of your house/company.
- Maintenance: Is interesting to have all the electric installation updated, and without imperfections. The maintenance subservience helps to achieve that.
- Insurance: Electric insurance, for any possible unexpected costs and/or risk at your house/enterprise.
- Green energy providing: Nowadays, people is environment aware, so for a lot of people is important to ensure that the energy that they use provides from green source as solar panels, windmills and any other renewable source.
- Business manager: It is a person that can check your bills, and any other requirement to propose improvements of your products.

Without forgetting the funding preferred terms of a customer for the sub products that will be easy paid with banking sector help.

This idea has been seen in the researches did by Salman Nazari-Shirkouhi et al. and Lutters et.al.

This sub services, are not a must, obviously, but they improve the service offered by the company and can reduce in long term the total bill import. Because this, the poll does not ask if you want that or not. It asks your interest against every subservice.
The possible answers have a range from 1 to 5, and later the algorithm applies a translation of this number to a percentage of interest. This interest is computed with a stochastic process. Every number has a minimum and a maximum probability, and then you take a random number between both.

This stochasticity gives to the algorithm the power to give different results for the same answers.
It is very interesting because, when a costumer says that they are very interested (5 points) to a product it does not mean that they will pay anything. This tells you that are interested, but the question is how really interested are they? With stochastic process we can achieve a value generated between what the company considers the minimum and maximum limit of this interest.

For example, a 5, obviously is not a maybe. It is I am highly interested but it is possible that I say no to the subservice because for me is very expensive. Or even, I do not fully understand what the sub product offers. Then, this 5 can be cast to any number between 80% and 100%.

This last paragraph gives the introduction to the next algorithm key point.
It has been told that the interest depends on what the subservice offers and the real price.
Probably if the green energy costs 1 euro, everybody, even the one that answers with a 1 in the survey will pay for it. But if it costs 1000 euros, even some of the customers that say 5, will avoid this sub product.

Then, it is important to estimate the maximum value that a customer is interested to pay for a specific subservice. With this we achieve two things:

1. Normalize subservices:
   Is not the same a subservice that costs 1 euro, that the one that costs 1000 euros and need funding during one full year. To achieve a fair result in the algorithm all sub services must be treated as similar, so it is important to normalize data.

2. Compute new optimization algorithm parameters:
   The optimization algorithm will be covered later with detail, but it will be the knapsack algorithm. This needs a list of tools that can be introduced into the knapsack, and the value of every tool, without forgetting the maximum capacity of it.

We have all parameters with business requirements and customer survey answer, but we need to compute the maximum capacity of every customer knapsack.

This is an interesting thing, every customer has a different “knapsack”, different availability to pay for a product, and this creates a lot of possibilities for the algorithm.
To do the first point the algorithm multiplies the stochastic interest percentage per the real subservice costs.

If a customer says a 4 for battery subservice, it is translated a number between 60% and 80%. Let supposes 75%. Then, if this cost 50 euro every month the customer availability to pay will be: 50€ * 0.75 = 37.5€.

When you apply this for all the sub products you have a list of real costs and expected costs for every customer. If they are sum, you have that a product with all sub products that fit the business requirements for this customer costs 500 euros, but the customer can pay 275 euros.

Both values will be parameters of the knapsack algorithm (point two) that will be seen later. For the moment, can be specified that 275 euros is the limit, but you have a lot of interesting combinations to achieve a product that does not past this maximum value.

1.3.2.3 Knapsack algorithm

Then, it is time for the product generation algorithm. This will try to create for each input customer a “perfect” product. To be able to do this the algorithm needs:

1. A new customer with the respective input data
2. Execute customer clustering
3. Compute the interest of a customer for every subservice.
4. A set of business restrictions: That allow to understand which level of subservice can be provided to the customer.
5. Maximum availability to pay:
   This is what really matters, it a principal key factor of economics. Everybody buy the product that the price corresponds to the customer value. It is related with customer happiness. If I pay more for a product that what I am likely to pay, I will search any other product that fits me.

Both first points are related with the explained above. Here the interesting point starts with the third point.

As explained previously, the algorithm will take all the answers of the questions related to customer interest and translate it to a stochastic interest percentage. Then, it will be multiplied per the real cost if this subservice. Every subservice has different levels, an easy explanation is that it is not the same the insurance of a house that an enterprise. This means that there are some business restrictions to offer what a customer can have (Point 4).

Once, the algorithm knows the maximum level of a subservice, it computes the real cost. Some of this subservice have a random factor to compute costs, it adds again stochasticity to the algorithm. This is due because some subservices depend of consumption, hired power, etc... Factors that are not always sure, and the costs of this factors itself depends of the market.
This consumption, hired power, tariff and other parameters are gathered by customer answers, but are also gathered through the cluster that the customer belongs.
Maybe a customer has consumption 123, but the cluster has 198. Obviously, if we take one value we will get one cost that does not correspond with the other. But these risks are part of the use of Machine Learning techniques that use statistics as a tool.

With the real cost, and the percentage, when multiply both we achieve the total price expected by the customer. The maximum is obviously the real price.
Finally, the algorithm sums all total expects prices. Creating a global expected price. This is done two time, one for customer data, one for cluster data (the one that the customer belongs).

Example

Subservice	Interest	Real cost	Expected customer cost
Electrical car	0,8	100	80
Smart home	0,2	80	16
Maintenance	0,6	10	6
Insurance	0,8	6	4,8
Green energy	0,9	0,83	0,747
Business manager	0,1	22	2,2
Total	**-**	**218,83**	**109,747**

Table 1: Table of subservices and their prices

Once all of this is calculated the algorithm start to iterate over all the possible combinations of subservices, contemplating the real cost, but using as a restriction the total expected customer cost.

Taking a quick look, some combinations that fulfil those requirements are:

- Smart home & Maintenance & Insurance & Green energy=96,83 €
- Electrical car & insurance & green energy=106,83 €
- Maintenance & Insurance & Green energy & Business manager=38,83 €

Once we have this, we could use a determined strategy to produce a single output but is interesting the point to present different solutions and that an expert chose the better for the customer itself. We do not have to forget that Artificial Intelligence is a tool, in this particular case, a tool to create the best possible product contemplating a set of business, customer, and market requirements that produces a high manual tasks highly sensitive to human errors. The aim is to automatize the process and then an electric engineer, expert in this area, and who do this job currently check the process and chose the final output.

With this we achieve a set of things:

- The contact to the customer can be done personally by an employee, improving ratio of response, and selling products easily.
- Reduce error in product generation process.
- Dedicate the engineers to the tasks that really matters, and are difficult to automatize or even do not interest. This will be possible because automatic process reduce study time, manual work, and achieve a better data normalization for the companies who applies this algorithms. Moreover, will add high knowledge of who they customers are and what they want. And will facilitate product cross selling as insurance, maintenance and others from the main sold product.

Then, the algorithm produces a CSV file with the cluster product data generation algorithm, and other with the customer product data generation. Keeping on with the example above, the Output will be like it:

1. The cheapest product: Maintenance & Insurance & Green energy & Business manager=38,83 €
2. The expensive product: Electrical car & insurance & green energy=106,83 €
3. The best ratio product: Maintenance & Insurance & Green energy & Business manager=38,83 €

The first output means that the cheapest is probably what the engineer should offer if the customer is reticent to pay a lot for energy services.

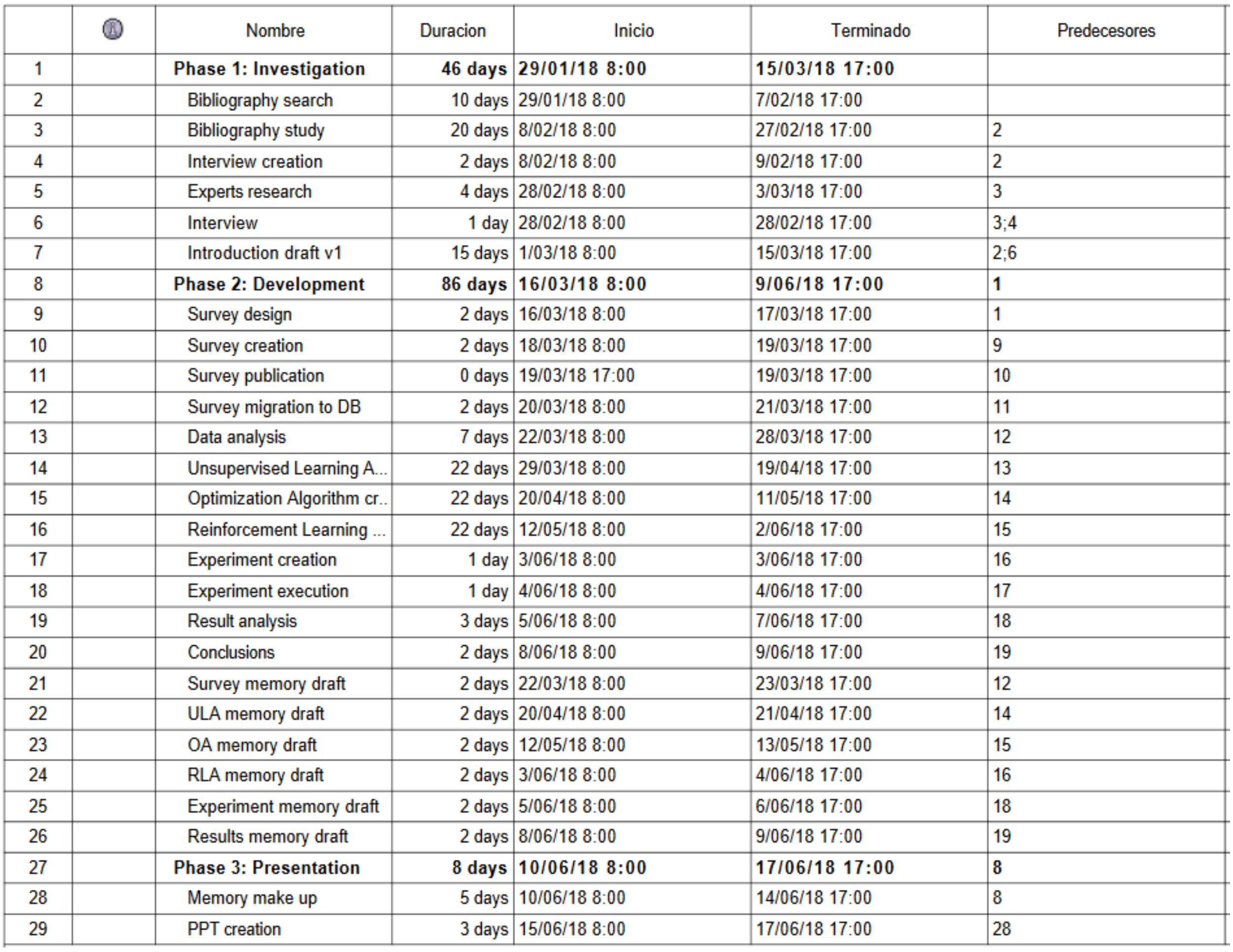
The second one is what interest more to the company, and finally the third (in this case is the same as the first) is what offers the best ratio total price for what it includes.

This output then is the best output to help the engineer, or the commercial to prepare the speech to sell the product to the costumer. First using what interest more to the company, but finally what interest more to the customer. Is important to consider that a customer that is in our company, is a costumer that does not pay for the same service to the competence.

1.4 Planning of the Work

To ensure that the project is delivered in the expected deadline, and to verify the constant development of work is useful to take advantage from Project Management tools.

In this case, it has been developed a Gantt diagram with the tasks to develop with the open-source software "Project Libre".
The tasks and their timing are:



	ⓘ	Nombre	Duracion	Inicio	Terminado	Predecesores
1		**Phase 1: Investigation**	**46 days**	**29/01/18 8:00**	**15/03/18 17:00**	
2		Bibliography search	10 days	29/01/18 8:00	7/02/18 17:00	
3		Bibliography study	20 days	8/02/18 8:00	27/02/18 17:00	2
4		Interview creation	2 days	8/02/18 8:00	9/02/18 17:00	2
5		Experts research	4 days	28/02/18 8:00	3/03/18 17:00	3
6		Interview	1 day	28/02/18 8:00	28/02/18 17:00	3;4
7		Introduction draft v1	15 days	1/03/18 8:00	15/03/18 17:00	2;6
8		**Phase 2: Development**	**86 days**	**16/03/18 8:00**	**9/06/18 17:00**	**1**
9		Survey design	2 days	16/03/18 8:00	17/03/18 17:00	1
10		Survey creation	2 days	18/03/18 8:00	19/03/18 17:00	9
11		Survey publication	0 days	19/03/18 17:00	19/03/18 17:00	10
12		Survey migration to DB	2 days	20/03/18 8:00	21/03/18 17:00	11
13		Data analysis	7 days	22/03/18 8:00	28/03/18 17:00	12
14		Unsupervised Learning A...	22 days	29/03/18 8:00	19/04/18 17:00	13
15		Optimization Algorithm cr...	22 days	20/04/18 8:00	11/05/18 17:00	14
16		Reinforcement Learning ...	22 days	12/05/18 8:00	2/06/18 17:00	15
17		Experiment creation	1 day	3/06/18 8:00	3/06/18 17:00	16
18		Experiment execution	1 day	4/06/18 8:00	4/06/18 17:00	17
19		Result analysis	3 days	5/06/18 8:00	7/06/18 17:00	18
20		Conclusions	2 days	8/06/18 8:00	9/06/18 17:00	19
21		Survey memory draft	2 days	22/03/18 8:00	23/03/18 17:00	12
22		ULA memory draft	2 days	20/04/18 8:00	21/04/18 17:00	14
23		OA memory draft	2 days	12/05/18 8:00	13/05/18 17:00	15
24		RLA memory draft	2 days	3/06/18 8:00	4/06/18 17:00	16
25		Experiment memory draft	2 days	5/06/18 8:00	6/06/18 17:00	18
26		Results memory draft	2 days	8/06/18 8:00	9/06/18 17:00	19
27		**Phase 3: Presentation**	**8 days**	**10/06/18 8:00**	**17/06/18 17:00**	**8**
28		Memory make up	5 days	10/06/18 8:00	14/06/18 17:00	8
29		PPT creation	3 days	15/06/18 8:00	17/06/18 17:00	28

Table 2: List of project tasks

The gantt diagram:

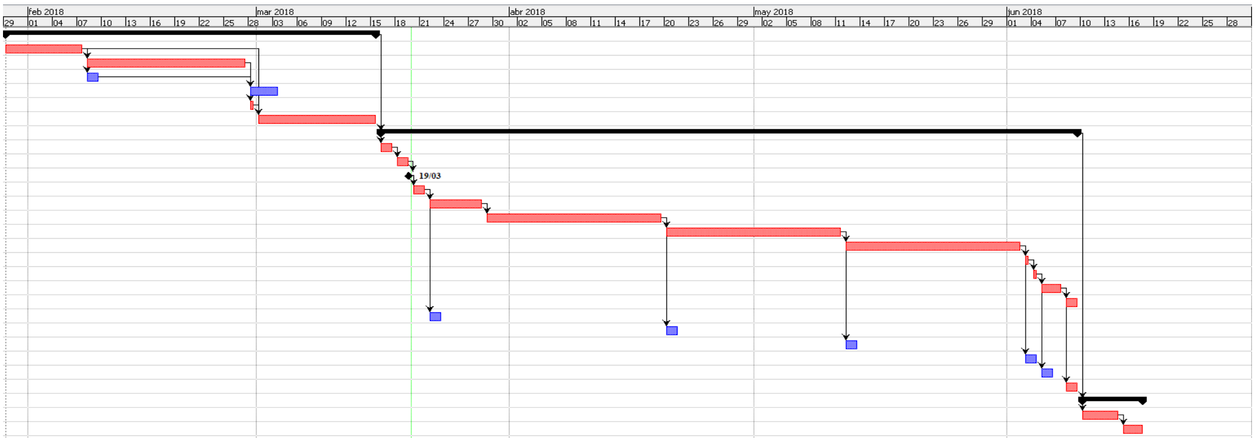


Figure 5: Gantt diagram

In the annexes you can find a pdf file with the project management report.

1.5 Brief summary of products obtained

This project will generate different products, but all of them will be highly valuable actives for today information technologies enterprises.

First, there is the algorithm. It will provide a Python code which applies clustering techniques and supervised learning techniques to classify customers and create the best fitting products for them based in a heuristics methodology.

Another great active is the database.
Despite this algorithm will be developed using invented data, the expertise in the energy sector of the developer and the interviewed agents will ensure the adaptability of this into the organizations.

Furthermore, the model will be clear and clean, and normalized to ensure the correct understanding of the variables and avoid redundant data. This will provide a highly understandable database but small enough to quickly introduce into companies.

Finally, the memory itself will be the last generated product. It provides how the work has been done. It helps another researcher/developer to reproduce the work and add their improvements, contributing this way into the open source software that is so important nowadays.

1.6 Brief description of the other chapters of the memory

In the following chapters, the present document will establish some interesting points for the project.

First, it will introduce a conceptual framework where will be studied other similar project in sector, techniques or any other variable that makes such previous researches interesting in their application. To achieve that the bibliography will be reviewed.

Then, derivate from this theoretical knowledge will be explained the methodology. This chapter will include a high level of explanation. There is possible to find programming languages, version of software, flow diagrams and others that create a general overview for the problem.
It also includes inputs, outputs, pseudo code and others.

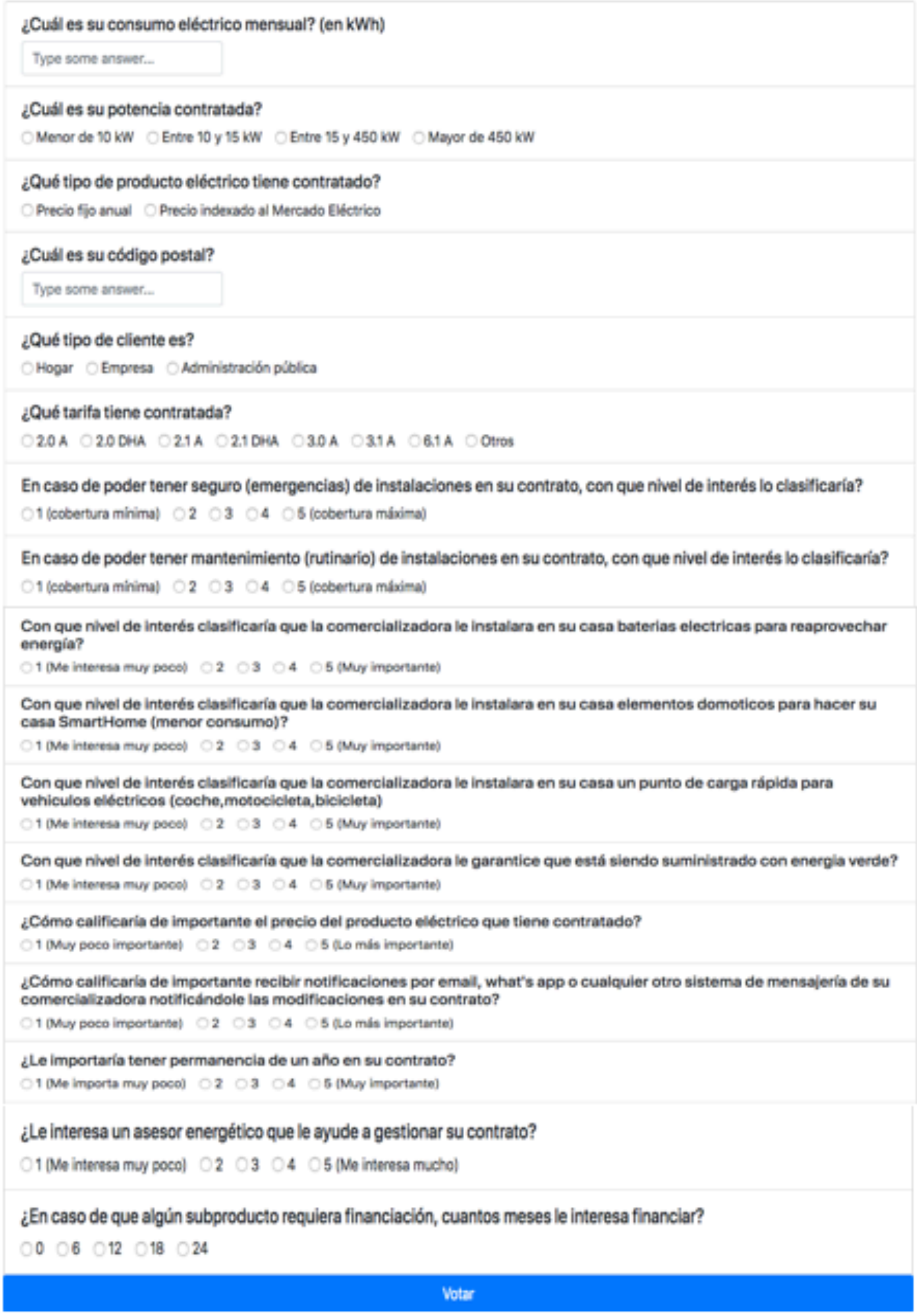
The last development chapter will explain how testing will be implemented and how will the model be reefed to ensure the supervised learning. Furthermore, is possible to find the results obtained through this testing actions.

Finally, these results will be summarized and analysed in the conclusions chapter.

The last lines of this project will be the bibliography and the attached documents.

2. Data gathering design

Afterwards literature review is easy to understand that a project key is to achieve the correct data to be able to ensure the correct clustering of customers, and subsequently create the most fitting product for them.



¿Cuál es su consumo eléctrico mensual? (en kWh)

Type some answer...

¿Cuál es su potencia contratada?

○ Menor de 10 kW ○ Entre 10 y 15 kW ○ Entre 15 y 450 kW ○ Mayor de 450 kW

¿Qué tipo de producto eléctrico tiene contratado?

○ Precio fijo anual ○ Precio indexado al Mercado Eléctrico

¿Cuál es su código postal?

Type some answer...

¿Qué tipo de cliente es?

○ Hogar ○ Empresa ○ Administración pública

¿Qué tarifa tiene contratada?

○ 2.0 A ○ 2.0 DHA ○ 2.1 A ○ 2.1 DHA ○ 3.0 A ○ 3.1 A ○ 6.1 A ○ Otros

En caso de poder tener seguro (emergencias) de instalaciones en su contrato, con que nivel de interés lo clasificaría?

○ 1 (cobertura mínima) ○ 2 ○ 3 ○ 4 ○ 5 (cobertura máxima)

En caso de poder tener mantenimiento (rutinario) de instalaciones en su contrato, con que nivel de interés lo clasificaría?

○ 1 (cobertura mínima) ○ 2 ○ 3 ○ 4 ○ 5 (cobertura máxima)

Con que nivel de interés clasificaría que la comercializadora le instalara en su casa baterias electricas para reaprovechar energía?

○ 1 (Me interesa muy poco) ○ 2 ○ 3 ○ 4 ○ 5 (Muy importante)

Con que nivel de interés clasificaría que la comercializadora le instalara en su casa elementos domoticos para hacer su casa SmartHome (menor consumo)?

○ 1 (Me interesa muy poco) ○ 2 ○ 3 ○ 4 ○ 5 (Muy importante)

Con que nivel de interés clasificaría que la comercializadora le instalara en su casa un punto de carga rápida para vehiculos eléctricos (coche,motocicleta,bicicleta)

○ 1 (Me interesa muy poco) ○ 2 ○ 3 ○ 4 ○ 5 (Muy importante)

Con que nivel de interés clasificaría que la comercializadora le garantice que está siendo suministrado con energia verde?

○ 1 (Me interesa muy poco) ○ 2 ○ 3 ○ 4 ○ 5 (Muy importante)

¿Cómo calificaría de importante el precio del producto eléctrico que tiene contratado?

○ 1 (Muy poco importante) ○ 2 ○ 3 ○ 4 ○ 5 (Lo más importante)

¿Cómo calificaría de importante recibir notificaciones por email, what's app o cualquier otro sistema de mensajería de su comercializadora notificándole las modificaciones en su contrato?

○ 1 (Muy poco importante) ○ 2 ○ 3 ○ 4 ○ 5 (Lo más importante)

¿Le importaría tener permanencia de un año en su contrato?

○ 1 (Me importa muy poco) ○ 2 ○ 3 ○ 4 ○ 5 (Muy importante)

¿Le interesa un asesor energético que le ayude a gestionar su contrato?

○ 1 (Me interesa muy poco) ○ 2 ○ 3 ○ 4 ○ 5 (Me interesa mucho)

¿En caso de que algún subproducto requiera financiación, cuantos meses le interesa financiar?

○ 0 ○ 6 ○ 12 ○ 18 ○ 24

Figure 6: Customer poll

2.1 Customer data survey

Has been told that the project must be an online tool. This is to achieve the purpose of constant product generation for any customer. For example, the tool can be included in the company website to achieve interesting leads.
Then, the landing page is a survey with a few questions that has been consciously design.

2.2 Customer product data survey

The poll design has been done by the project researcher with the help of industrial and electrical engineers that nowadays develop their tasks on different marketers between United Kingdom and Spain. To achieve deeper understanding of design methodology is important to read the attached interview in this project.
With the answers provided by the consultant and the two-year experience in the electrical sector of the researcher a 360 vision has been achieved of the business and it requirements to create the complete data pipeline of data gathering.

As in any sector, a customer can be defined with some variables, in the electrical sector what defines a customer are the following points:

- Electrical consumption (Monthly)
- Hired power
- Customer type
- Tariff

With these four features, that also between them are completely related, we can understand what we can expect of each customer and move on what can we offer to them.

2.3 Customer most valuable product features survey

Once, we know what a customer is, and we are able to classify them, the goal is to understand what makes happy a customer. This in marketing can be understood of what subservices of a product presents extra value and what do not.

An example using telecommunications sector is the offer of a payment subscription to some streaming television service such as Netflix or HBO. A lot of customers evaluate it positively. If you ask a customer does it makes you happy? The answer in a lot of cases will be yes.

Despite that cannot be forgotten that this project inherits marketing developments, such as surveys design, and the communication with the customer must be clear, precise and not ambiguous. Then, in the poll it is not asked if a subservice makes you happy. What it asks if this subservice have value for you, or in other words, using points, ask for what do really interest you.

Using the previous example, the correct question is: Do Netflix/HBO really interests you in the television service?
The answer can be any value between 1 and 5, where 5 is Yes, I really do. And 1 means it does not interest me.

Using this the poll achieve the clear, and precise questions that we should aim as some bibliography recommends, but also closed questions. Data analysis depends on data achieved by data engineers. They depend directly of what software developers allow to a customer in their websites and software's. If the website allows a full open text box, the data analysis can have data with no value. Not for what customer writes, but for IT limitations. Is easier for an algorithm understand a value between 1 and 5 that in the data pipeline is translated to numeric values, that a text that says: "Yes, I love Netflix because they have "Narcos" TV show.". This example shows that a lot of data is unprofitable due to his nature. To analyse this should be applied a well-developed Natural Language Processing Algorithm (NLP).

Considering all this stuff is feasible to achieve analysable data with customer real answers that provide value of product features that would like to be analysed.

2.4 New product happiness survey

The initial idea for the project was to develop different questions in the survey.

The first brainstorming provides that the survey should ask the customer for what they are, this means the four big features that allow to classify them.

A second turn of questions that asks for what they currently have, and then the last turn for what really interest them, what makes them happy.

Later, when the whole survey was developed, the study did for the researcher and the consultants that provides business knowledge into the project saw that such a long survey is not practical, any customer would answer it and neither will provide extra value because the correlation between what they have and what makes them happy.

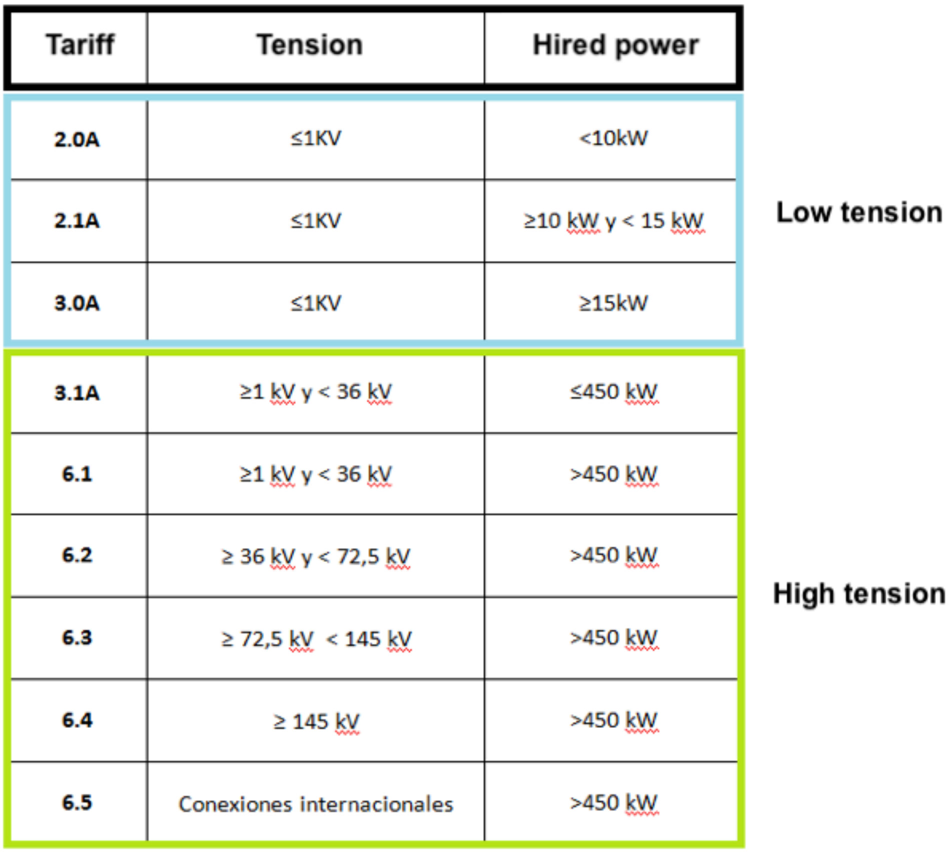
If one feature that they have does not make them happy it does not make sense to evaluate why they have it, because this is more related of what the current company offers them, that customer preferences. Despite, it could be about customer preferences makes references of past preferences, which in time changes.

Then, this phase of new product happiness survey has been erased, and the questions have been checked a lot of times to only keep alive what really matters, and make sense for the projects, including User Usability and User Experience (UX).

3. Business rules gathering

Obviously, any algorithm, and more precisely an optimization algorithm has some restrictions that are related with business scope.

In this project all the business restrictions are related with what a customer is. This means that the features that describe a customer and are the essence of the clustering algorithm that will be seen later are related with the limits of what a company can offer to the specific user.



Tariff	Tension	Hired power
2.0A	≤1KV	<10kW
2.1A	≤1KV	≥10 kW y < 15 kW
3.0A	≤1KV	≥15kW
3.1A	≥1 kV y < 36 kV	≤450 kW
6.1	≥1 kV y < 36 kV	>450 kW
6.2	≥ 36 kV y < 72,5 kV	>450 kW
6.3	≥ 72,5 kV < 145 kV	>450 kW
6.4	≥ 145 kV	>450 kW
6.5	Conexiones internacionales	>450 kW

Low tension

High tension

Table 3: Tariff and hired power equivalence table

The above table shows that tariff and hired power is directly related. Then using the tariff, we can calculate the estimated cost of KWh and any other unit of any subservice and the realizing the corresponding calculations with the energy consumption we have expected costs.

These costs are the following:

- **Access tariff cost:** This cost is related with tariff, hired power and consumption.
  - **TE:** Energy term = €/kWh
  - **TP:** Power term = €/kWh*year
  - **Cost 1:** Tp * hired_power
  - **Cost 2:** (Te + other costs) * consumption

- **Pure electricity cost:** Cost related with tariff that the customer pays for energy supply.

- **OMIE electric network cost:** Network maintenance cost
  - Fix cost: 0,0006 €/kWh (BOE)

- **Pool energy cost:** What the marketer pays for buy the energy that sell to the customers. It has variance per day and is calculated by an intermediary agent called OMIE.

- **Deviation cost:** When a marketer buys energy it buys packs of this energy. When they buy more than what they really sell it have an extra cost, because the marketer is booking or blocking more energy what they consume, then the market has to generate it to provide and nobody use it. The same happens when the company have to buy more energy later the market is close. An extra cost is paid due to the forecasting mistake.

- **Power warranty cost:** A cost fixed for tariff.

- **Governance fees:**
  - Electrical fee
  - Municipal fee (5%)
  - VAT

- **Green energy cost**: Fixed cost of 0,83€/month

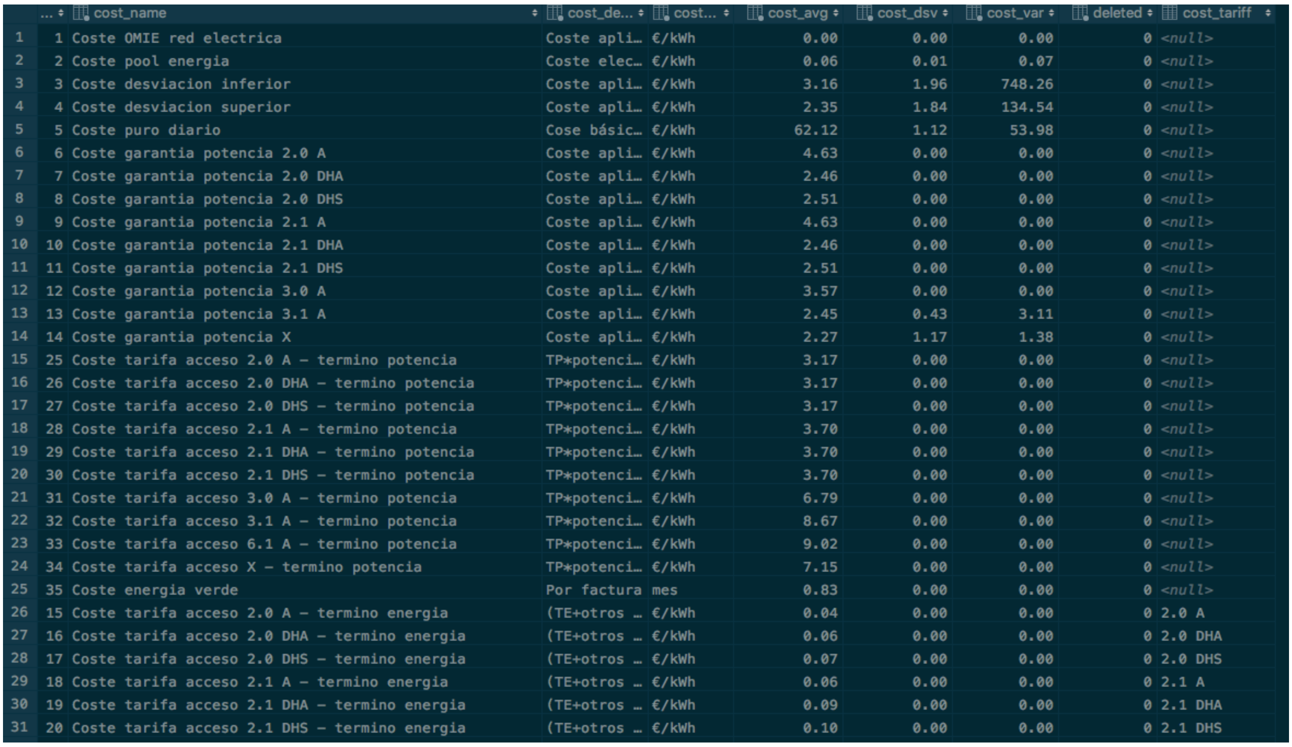
A big deal is to consider that energy supply is without leaks. In a real case environment, the marketer has the leak ratio and then has to add it in his calculations to buy the real energy.

All these costs are regulated by OMIE, or the governance. Both are intermediaries in the current public energy sector in Spain.

At website bibliography is possible to find all the links that provide the electric costs. Within this links exists some download links from zip files that includes historical data of prices. For this project the researcher took the minimum and maximum cost of July and add random calculation between these values in the product generation algorithm (it will be seen later).

At attached documents exists “cost_matrix.xlsx” with computes the average value of every single day for the costs that does not present variation between hours.

This tables have been translated into SQL table that has computes the average cost and the deviation and variation of these costs. The aim is to fully represents the market variability between days and hours.



	...	cost_name	cost_de...	cost...	cost_avg	cost_dsv	cost_var	deleted	cost_tariff
1	1	Coste OMIE red electrica	Coste apli…	€/kWh	0.00	0.00	0.00	0	<null>
2	2	Coste pool energia	Coste elec…	€/kWh	0.06	0.01	0.07	0	<null>
3	3	Coste desviacion inferior	Coste apli…	€/kWh	3.16	1.96	748.26	0	<null>
4	4	Coste desviacion superior	Coste apli…	€/kWh	2.35	1.84	134.54	0	<null>
5	5	Coste puro diario	Cose básic…	€/kWh	62.12	1.12	53.98	0	<null>
6	6	Coste garantia potencia 2.0 A	Coste apli…	€/kWh	4.63	0.00	0.00	0	<null>
7	7	Coste garantia potencia 2.0 DHA	Coste apli…	€/kWh	2.46	0.00	0.00	0	<null>
8	8	Coste garantia potencia 2.0 DHS	Coste apli…	€/kWh	2.51	0.00	0.00	0	<null>
9	9	Coste garantia potencia 2.1 A	Coste apli…	€/kWh	4.63	0.00	0.00	0	<null>
10	10	Coste garantia potencia 2.1 DHA	Coste apli…	€/kWh	2.46	0.00	0.00	0	<null>
11	11	Coste garantia potencia 2.1 DHS	Coste apli…	€/kWh	2.51	0.00	0.00	0	<null>
12	12	Coste garantia potencia 3.0 A	Coste apli…	€/kWh	3.57	0.00	0.00	0	<null>
13	13	Coste garantia potencia 3.1 A	Coste apli…	€/kWh	2.45	0.43	3.11	0	<null>
14	14	Coste garantia potencia X	Coste apli…	€/kWh	2.27	1.17	1.38	0	<null>
15	25	Coste tarifa acceso 2.0 A – termino potencia	TP*potenci…	€/kWh	3.17	0.00	0.00	0	<null>
16	26	Coste tarifa acceso 2.0 DHA – termino potencia	TP*potenci…	€/kWh	3.17	0.00	0.00	0	<null>
17	27	Coste tarifa acceso 2.0 DHS – termino potencia	TP*potenci…	€/kWh	3.17	0.00	0.00	0	<null>
18	28	Coste tarifa acceso 2.1 A – termino potencia	TP*potenci…	€/kWh	3.70	0.00	0.00	0	<null>
19	29	Coste tarifa acceso 2.1 DHA – termino potencia	TP*potenci…	€/kWh	3.70	0.00	0.00	0	<null>
20	30	Coste tarifa acceso 2.1 DHS – termino potencia	TP*potenci…	€/kWh	3.70	0.00	0.00	0	<null>
21	31	Coste tarifa acceso 3.0 A – termino potencia	TP*potenci…	€/kWh	6.79	0.00	0.00	0	<null>
22	32	Coste tarifa acceso 3.1 A – termino potencia	TP*potenci…	€/kWh	8.67	0.00	0.00	0	<null>
23	33	Coste tarifa acceso 6.1 A – termino potencia	TP*potenci…	€/kWh	9.02	0.00	0.00	0	<null>
24	34	Coste tarifa acceso X – termino potencia	TP*potenci…	€/kWh	7.15	0.00	0.00	0	<null>
25	35	Coste energia verde	Por factura	mes	0.83	0.00	0.00	0	<null>
26	15	Coste tarifa acceso 2.0 A – termino energia	(TE+otros …	€/kWh	0.04	0.00	0.00	0	2.0 A
27	16	Coste tarifa acceso 2.0 DHA – termino energia	(TE+otros …	€/kWh	0.06	0.00	0.00	0	2.0 DHA
28	17	Coste tarifa acceso 2.0 DHS – termino energia	(TE+otros …	€/kWh	0.07	0.00	0.00	0	2.0 DHS
29	18	Coste tarifa acceso 2.1 A – termino energia	(TE+otros …	€/kWh	0.06	0.00	0.00	0	2.1 A
30	19	Coste tarifa acceso 2.1 DHA – termino energia	(TE+otros …	€/kWh	0.09	0.00	0.00	0	2.1 DHA
31	20	Coste tarifa acceso 2.1 DHS – termino energia	(TE+otros …	€/kWh	0.10	0.00	0.00	0	2.1 DHS

Table 4: Cost table (database

This is a business requirement, if the algorithm has been developed in a deterministic methodology the application of it would be very limited, so it does not add value to the company and then the whole project lose sense.

Despite that, the business restrictions are very small. This is because a marketer can offer what they consider want. The only requirement of products is to compute correctly all the inherits costs and do not forget anything.

Another requirement is the subservice list. In this case the list of subservices uses a mix of what normal marketers offers in Spain and new products that are only able in foreign countries like Scotland.

The list is:

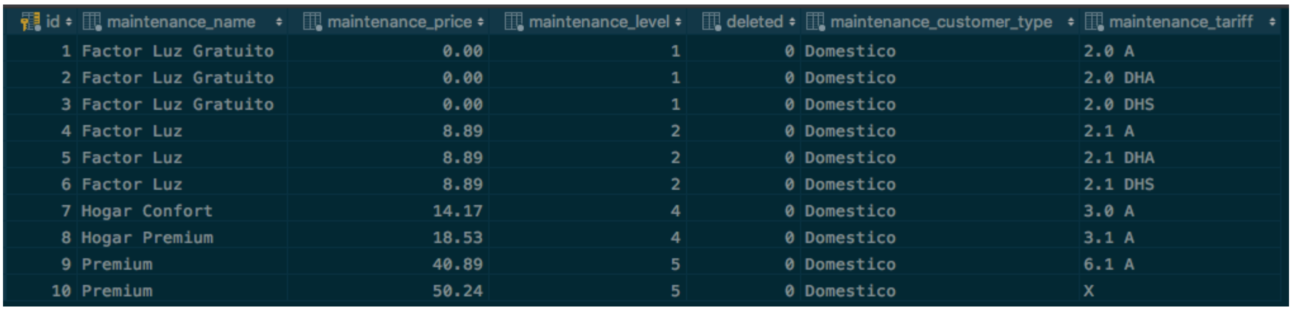
- Electrical car: This is very limited yet.
- Battery
- Smart home: This is only able in foreign countries.
- Maintenance
- Insurance
- Green energy providing
- Business manager

This set of subservices makes the algorithm complete and creates a great number of combinations and possibilities for the algorithm considering what they offer.

Finally, the last requirement in the algorithm is to develop a methodology that limits the product generation in what offers value to the customer. The companies are really worried about automatic process that can destroy the image, the engagement that they have generated with customers after years of investment and correct actions.

This makes that the algorithm will not be fully automatic, and the output should be provided to engineers that check what is the generated product. This generated product must cost less of what an engineer can imagine. This means that for example a small customer, tariff 2.0 A cannot receive an offer with the full premium maintenance that costs 200 euro monthly focused on enterprises.

This has been modelled in the data model with the field tariff in every table, for example maintenances table:



id	maintenance_name	maintenance_price	maintenance_level	deleted	maintenance_customer_type	maintenance_tariff
1	Factor Luz Gratuito	0.00	1	0	Domestico	2.0 A
2	Factor Luz Gratuito	0.00	1	0	Domestico	2.0 DHA
3	Factor Luz Gratuito	0.00	1	0	Domestico	2.0 DHS
4	Factor Luz	8.89	2	0	Domestico	2.1 A
5	Factor Luz	8.89	2	0	Domestico	2.1 DHA
6	Factor Luz	8.89	2	0	Domestico	2.1 DHS
7	Hogar Confort	14.17	4	0	Domestico	3.0 A
8	Hogar Premium	18.53	4	0	Domestico	3.1 A
9	Premium	40.89	5	0	Domestico	6.1 A
10	Premium	50.24	5	0	Domestico	X

Table 5: Example of subservice database table (maintenances)

This simple idea in data modelling allows to ensure that business requirements is correct.

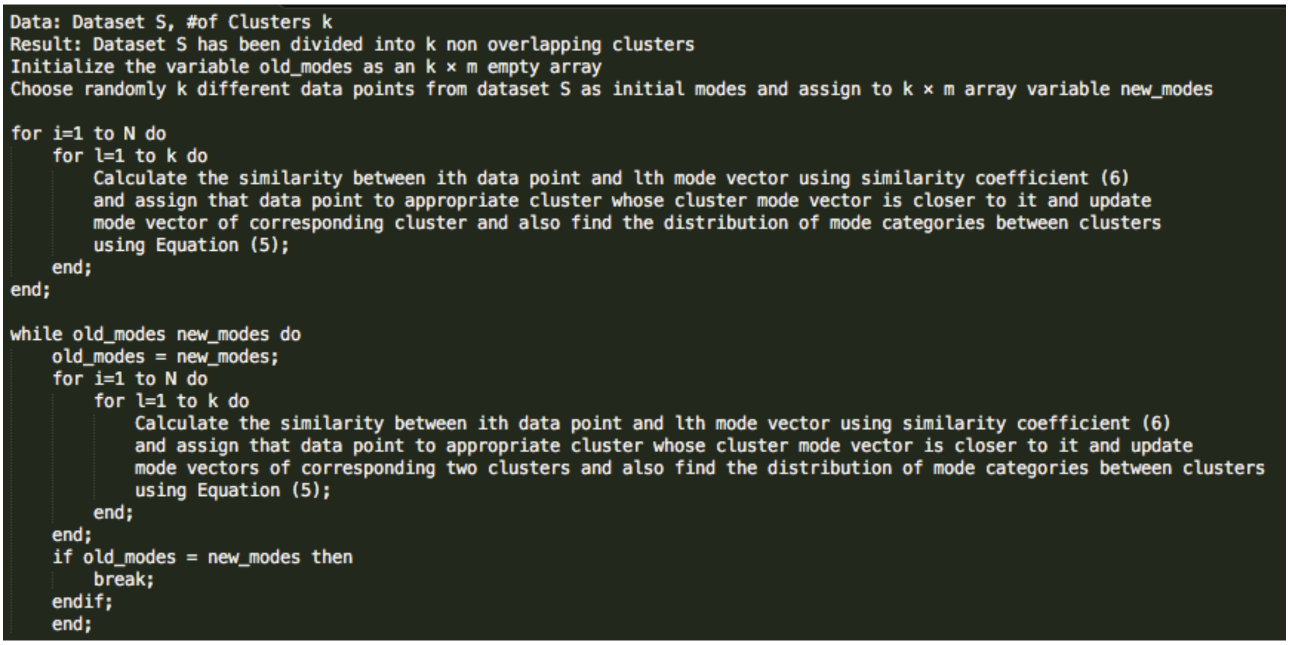
These whole business requirements have been done with the consultants, and with some work mates of the researcher in a medium marketer in Spain. The process has been very clear because despite the sources of the information were in different countries, the core of products remains the same for all of them, then the requirements were identical or very similar.

Related with this the survey provides the interest for each subservice, then with this the algorithm can compute the availability of payment of a customer to some subservice in the range of subservices for their tariff. This limits the offer of products and the maximum desired cost of a sub product for a customer.

The main conclusion here is that an algorithm cannot offer to a customer anything that do not correspond to the main features of a customer. Neither can forget some component of costs.

4. Unsupervised Learning Algorithm

The unsupervised Learning Algorithm that has been applied is a mix between K-means clustering and k-modes algorithm, the name of this mix is K-Prototypes algorithm.
The data nature of this project is mixed. This means that some data types are numbers (integers), and others are strings.



```
Data: Dataset S, #of Clusters k
Result: Dataset S has been divided into k non overlapping clusters
Initialize the variable old_modes as an k × m empty array
Choose randomly k different data points from dataset S as initial modes and assign to k × m array variable new_modes

for i=1 to N do
    for l=1 to k do
        Calculate the similarity between ith data point and lth mode vector using similarity coefficient (6)
        and assign that data point to appropriate cluster whose cluster mode vector is closer to it and update
        mode vector of corresponding cluster and also find the distribution of mode categories between clusters
        using Equation (5);
    end;
end;

while old_modes new_modes do
    old_modes = new_modes;
    for i=1 to N do
        for l=1 to k do
            Calculate the similarity between ith data point and lth mode vector using similarity coefficient (6)
            and assign that data point to appropriate cluster whose cluster mode vector is closer to it and update
            mode vectors of corresponding two clusters and also find the distribution of mode categories between clusters
            using Equation (5);
        end;
    end;
    if old_modes = new_modes then
        break;
    endif;
    end;
```

Figure 7: K-Modes pseudo code

k-means classify a given data set through a certain number of clusters fixed apriori. The idea is to define k centres. These centres should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centre.

The pseudo code for this algorithm is:

1) Randomly select ‘c’ cluster centres.
2) Calculate the distance between each data point and cluster centres.
3) Assign the data point to the cluster centre whose distance from the cluster centre is minimum of all the cluster centres.
4) Recalculate the new cluster centre
5) Recalculate the distance between each data point and new obtained cluster centres.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

Then, it is important to understand that this algorithm works with distances, and the only natural data set that have distance is numbers. For example, if a customer has 1000 KWh as consumption, another 2200 KWh and another one has 3000 KWh, and the algorithm is set with 2 clusters. The middle customer will be classified in the cluster near to 3000 KWh.

This does not make sense in data that are strings, for example tariff. To achieve a better understanding is good to think about: What is the difference between tariff 2.0 A and 2.0 DHA?

To answer this question, the fact is that k-means do not help, and then it is important to use a proper algorithm to cluster text data sets. This algorithm is k-modes it is used for clustering categorical variables. It defines clusters based on the number of matching categories between data points.

Both pseudo codes explain the algorithm developed in scikit-learn, and the specific version improved by Nicodb at GitHub https://github.com/nicodv/kmodes.
This project uses open source software to implement a full tested and optimized version of K-Prototype algorithm.

K-Prototypes in this case classify the following data:

- Electrical consumption (Monthly)
- Hired power
- Customer type
- Tariff

And depends on “form_token” id, which is the customer identifier that has replied to the survey. This data is obtained two times, one for the training data set, and another one for the data set to be predicted:

Training dataset:

```
SELECT * FROM crosstab($$ select form_token,question_id,answer_text from surveys_answer
where question_id in (1,2,3,5,6) and training_set=1 order by 1,2$$) AS final_result (submit_id
varchar(200),question1 varchar(200),question2 varchar(200), question3 varchar(200),question5
varchar(200),question6 varchar(200))
```

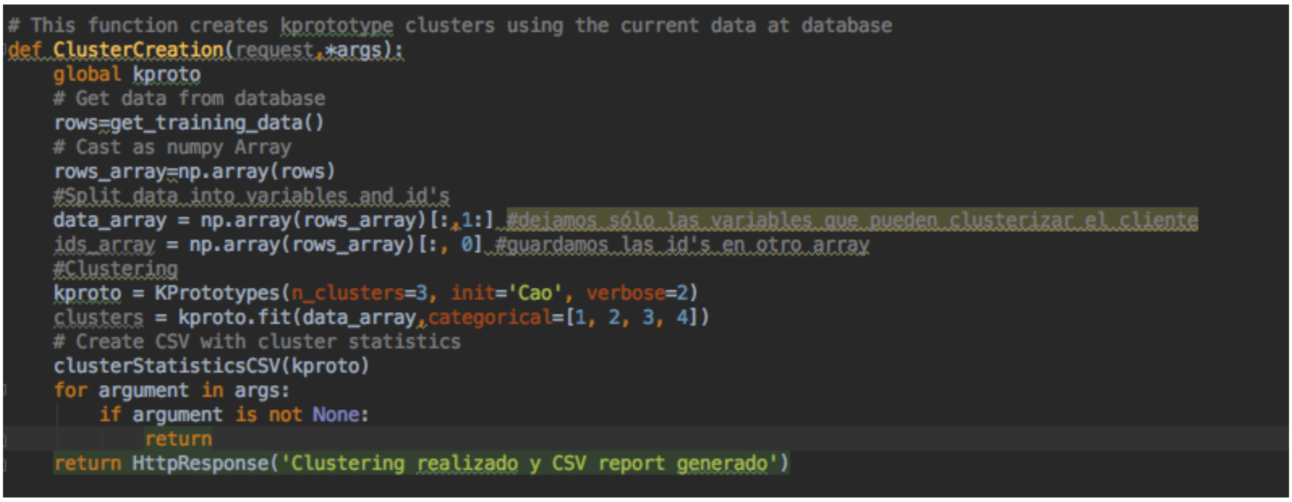
This dataset is the set of surveys that has been dedicated to cluster training. So, they define the different clusters.

To predict dataset:

```
SELECT * FROM crosstab($$ select form_token,question_id,answer_text from surveys_answer
where question_id in (1,2,3,5,6) and training_set=0 order by 1,2$$) AS final_result (submit_id
varchar(200),question1 varchar(200),question2 varchar(200), question3 varchar(200),question5
varchar(200),question6 varchar(200))
```

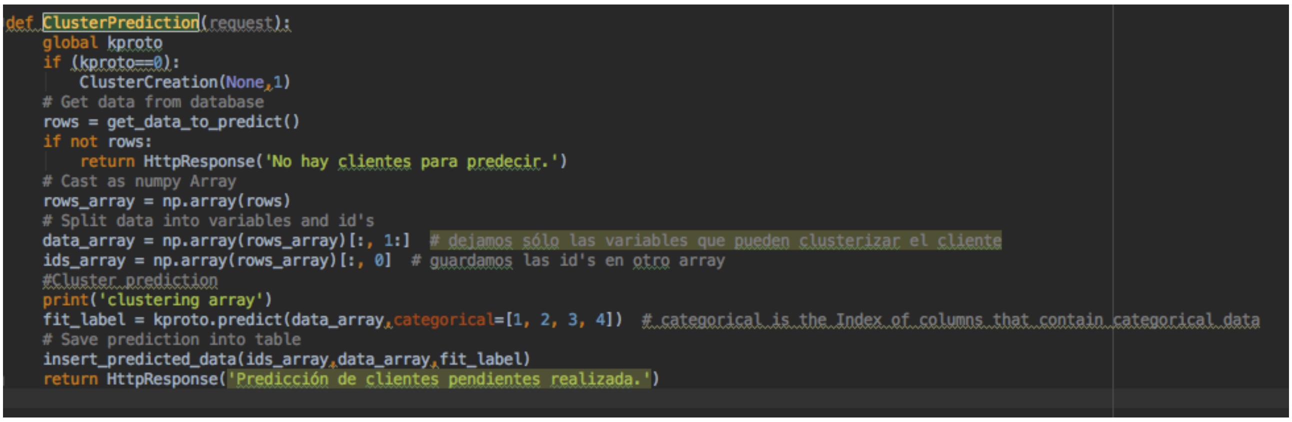
The second question (“question_id”=2) refers to consumption, which is a number data type, but the others makes references to text datasets. Then, K prototype will mostly use k-mode, but also k-means to calculate the Euclidean distance between consumptions.

This mix approach allows to provide a complete understanding of the variables that are part of a customer and then it is possible to cluster them. These lines of Python code do the clustering using the libraries explained before:



```
# This function creates kprototype clusters using the current data at database
def ClusterCreation(request,*args):
    global kproto
    # Get data from database
    rows=get_training_data()
    # Cast as numpy Array
    rows_array=np.array(rows)
    #Split data into variables and id's
    data_array = np.array(rows_array)[:,1:] #dejamos sólo las variables que pueden clusterizar el cliente
    ids_array = np.array(rows_array)[:, 0] #guardamos las id's en otro array
    #Clustering
    kproto = KPrototypes(n_clusters=3, init='Cao', verbose=2)
    clusters = kproto.fit(data_array,categorical=[1, 2, 3, 4])
    # Create CSV with cluster statistics
    clusterStatisticsCSV(kproto)
    for argument in args:
        if argument is not None:
            return
    return HttpResponse('Clustering realizado y CSV report generado')
```

Figure 8: ClusterCreation function



```
def ClusterPrediction(request):
    global kproto
    if (kproto==0):
        ClusterCreation(None,1)
    # Get data from database
    rows = get_data_to_predict()
    if not rows:
        return HttpResponse('No hay clientes para predecir.')
    # Cast as numpy Array
    rows_array = np.array(rows)
    # Split data into variables and id's
    data_array = np.array(rows_array)[:, 1:]  # dejamos sólo las variables que pueden clusterizar el cliente
    ids_array = np.array(rows_array)[:, 0]  # guardamos las id's en otro array
    #Cluster prediction
    print('clustering array')
    fit_label = kproto.predict(data_array,categorical=[1, 2, 3, 4])  # categorical is the Index of columns that contain categorical data
    # Save prediction into table
    insert_predicted_data(ids_array,data_array,fit_label)
    return HttpResponse('Predicción de clientes pendientes realizada.')
```

Figure 9: ClusterPrediction function

5. Product Generation Algorithm

The interesting part and what have been needed to develop by the researches is the product generation algorithm.
In previous sections is possible to understand what it will do, but not he full scope of it.

The starting point is that the customer has answered a survey. These answers are stored in a PostgreSQL database, and a K-Prototype algorithm is executed to know to which cluster the customer belongs. Once this happens, the next execution is the knapsack algorithm.

The knapsack problem or rucksack problem is a problem in *Combinatorial Optimisation*: Given a set of items, each with a weight and a value, determine the number of each item to include in a collection so that the total weight is less than or equal to a given limit and the total value is as large as possible. It derives its name from the problem faced by someone who is constrained by a fixed-size knapsack and must fill it with the most valuable items.

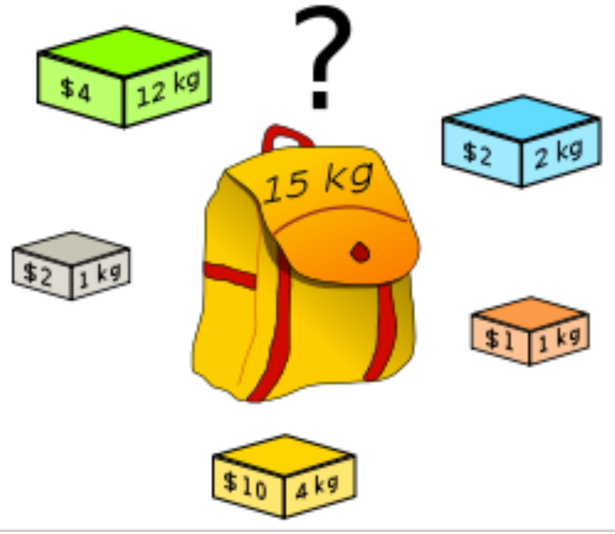
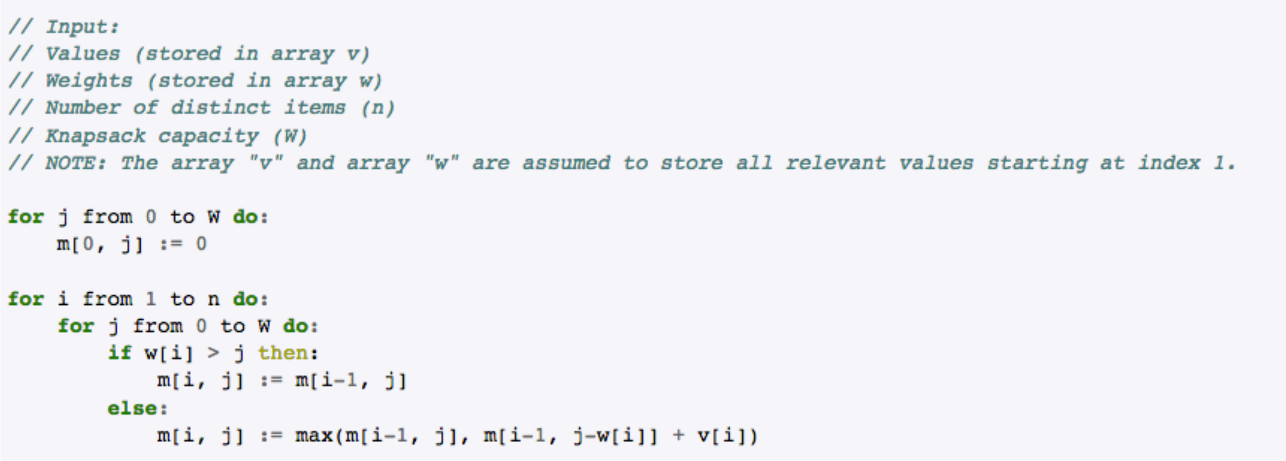


Figure 10: Knapsack problem schema

An example of pseudo code:



```
// Input:
// Values (stored in array v)
// Weights (stored in array w)
// Number of distinct items (n)
// Knapsack capacity (W)
// NOTE: The array "v" and array "w" are assumed to store all relevant values starting at index 1.

for j from 0 to W do:
    m[0, j] := 0

for i from 1 to n do:
    for j from 0 to W do:
        if w[i] > j then:
            m[i, j] := m[i-1, j]
        else:
            m[i, j] := max(m[i-1, j], m[i-1, j-w[i]] + v[i])
```

Figure 11: Knapsack problem pseudo code

Then, the important inputs are the “objects (values)” to store in knapsack, how much does an object weight, and finally the knapsack capacity.

It is possible to stablish similarity with the product generation problem.

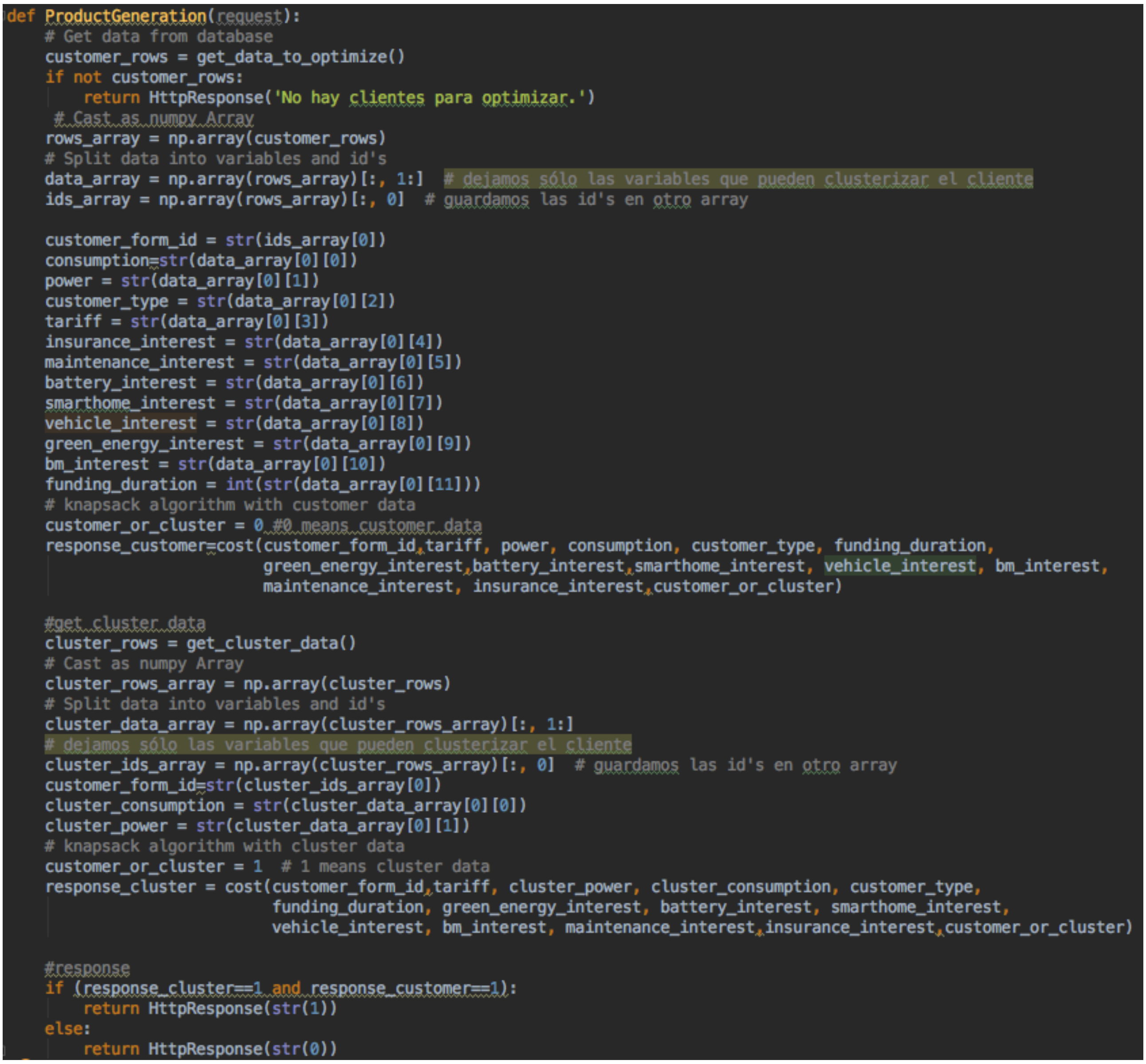
Objects: The company has a set of subservices that are part of the service that they offer.

Weight: The real cost of a subservice.

Capacity: The total amount that a customer is able or have willingness to pay.

Any customer will not fill their “product knapsack” with more products if the capacity is achieved. Then, another iteration of the fulfilling process must begin. If the algorithm stores the first feasible solution and try to sell it to the customer the algorithm will not take profit of the computational capacity, so it is important to compute the whole tree of feasible solutions and later choose which one is the “best”.

In the image below is possible to read the whole main program of the algorithm, but it calls a few functions that helps to compute the different product generation phases.



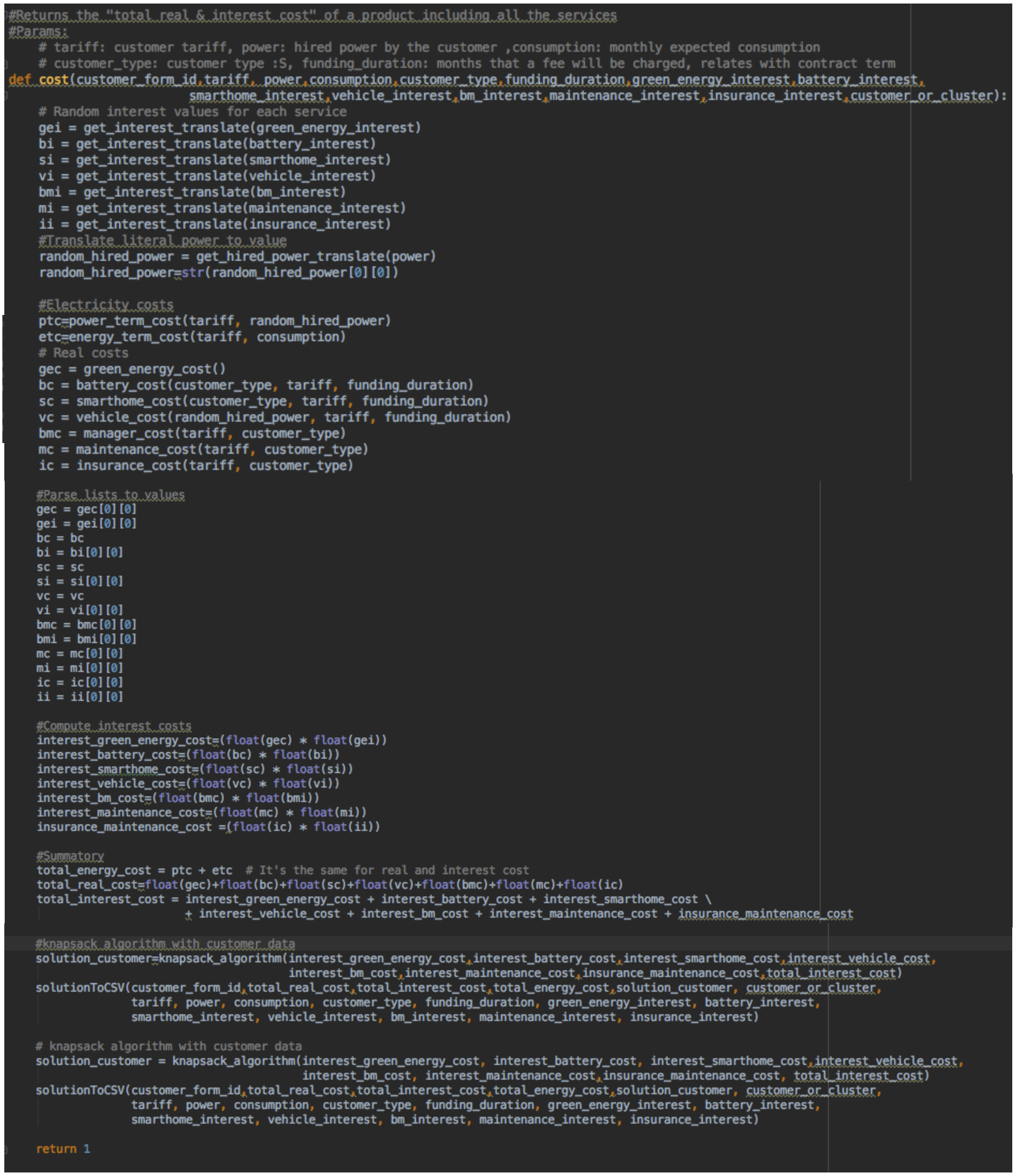
```
def ProductGeneration(request):
    # Get data from database
    customer_rows = get_data_to_optimize()
    if not customer_rows:
        return HttpResponse('No hay clientes para optimizar.')
     # Cast as numpy Array
    rows_array = np.array(customer_rows)
    # Split data into variables and id's
    data_array = np.array(rows_array)[:, 1:]  # dejamos sólo las variables que pueden clusterizar el cliente
    ids_array = np.array(rows_array)[:, 0]  # guardamos las id's en otro array

    customer_form_id = str(ids_array[0])
    consumption=str(data_array[0][0])
    power = str(data_array[0][1])
    customer_type = str(data_array[0][2])
    tariff = str(data_array[0][3])
    insurance_interest = str(data_array[0][4])
    maintenance_interest = str(data_array[0][5])
    battery_interest = str(data_array[0][6])
    smarthome_interest = str(data_array[0][7])
    vehicle_interest = str(data_array[0][8])
    green_energy_interest = str(data_array[0][9])
    bm_interest = str(data_array[0][10])
    funding_duration = int(str(data_array[0][11]))
    # knapsack algorithm with customer data
    customer_or_cluster = 0 #0 means customer data
    response_customer=cost(customer_form_id,tariff, power, consumption, customer_type, funding_duration,
                           green_energy_interest,battery_interest,smarthome_interest, vehicle_interest, bm_interest,
                           maintenance_interest, insurance_interest,customer_or_cluster)

    #get cluster data
    cluster_rows = get_cluster_data()
    # Cast as numpy Array
    cluster_rows_array = np.array(cluster_rows)
    # Split data into variables and id's
    cluster_data_array = np.array(cluster_rows_array)[:, 1:]
    # dejamos sólo las variables que pueden clusterizar el cliente
    cluster_ids_array = np.array(cluster_rows_array)[:, 0]  # guardamos las id's en otro array
    customer_form_id=str(cluster_ids_array[0])
    cluster_consumption = str(cluster_data_array[0][0])
    cluster_power = str(cluster_data_array[0][1])
    # knapsack algorithm with cluster data
    customer_or_cluster = 1  # 1 means cluster data
    response_cluster = cost(customer_form_id,tariff, cluster_power, cluster_consumption, customer_type,
                            funding_duration, green_energy_interest, battery_interest, smarthome_interest,
                            vehicle_interest, bm_interest, maintenance_interest,insurance_interest,customer_or_cluster)

    #response
    if (response_cluster==1 and response_customer==1):
        return HttpResponse(str(1))
    else:
        return HttpResponse(str(0))
```

Figure 12: Product generation function

The function cost of the algorithm takes all the customer data, including interest or all the cluster data, and the specific customer interest for the subservices and do the following (remember that this algorithm can run for the average, then it uses cluster data, or without the execution of the first algorithm and only uses customer feedback).

These functions help to compute the weight of each item and the total knapsack payment capacity.

The variable “total_interest_cost” stores the maximum cost that the full service can have for a customer.

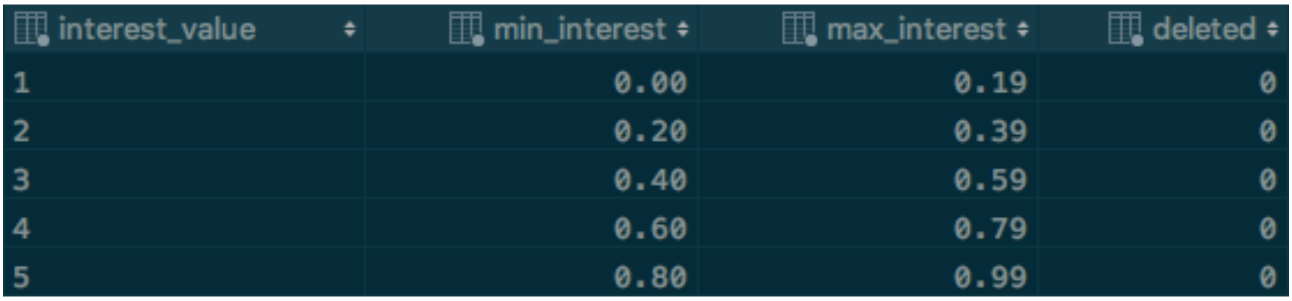


```python
#Returns the "total real & interest cost" of a product including all the services
#Params:
    # tariff: customer tariff, power: hired power by the customer ,consumption: monthly expected consumption
    # customer_type: customer type :S, funding_duration: months that a fee will be charged, relates with contract term
def cost(customer_form_id,tariff, power,consumption,customer_type,funding_duration,green_energy_interest,battery_interest,
                        smarthome_interest,vehicle_interest,bm_interest,maintenance_interest,insurance_interest,customer_or_cluster):
    # Random interest values for each service
    gei = get_interest_translate(green_energy_interest)
    bi = get_interest_translate(battery_interest)
    si = get_interest_translate(smarthome_interest)
    vi = get_interest_translate(vehicle_interest)
    bmi = get_interest_translate(bm_interest)
    mi = get_interest_translate(maintenance_interest)
    ii = get_interest_translate(insurance_interest)
    #Translate literal power to value
    random_hired_power = get_hired_power_translate(power)
    random_hired_power=str(random_hired_power[0][0])

    #Electricity costs
    ptc=power_term_cost(tariff, random_hired_power)
    etc=energy_term_cost(tariff, consumption)
    # Real costs
    gec = green_energy_cost()
    bc = battery_cost(customer_type, tariff, funding_duration)
    sc = smarthome_cost(customer_type, tariff, funding_duration)
    vc = vehicle_cost(random_hired_power, tariff, funding_duration)
    bmc = manager_cost(tariff, customer_type)
    mc = maintenance_cost(tariff, customer_type)
    ic = insurance_cost(tariff, customer_type)

    #Parse lists to values
    gec = gec[0][0]
    gei = gei[0][0]
    bc = bc
    bi = bi[0][0]
    sc = sc
    si = si[0][0]
    vc = vc
    vi = vi[0][0]
    bmc = bmc[0][0]
    bmi = bmi[0][0]
    mc = mc[0][0]
    mi = mi[0][0]
    ic = ic[0][0]
    ii = ii[0][0]

    #Compute interest costs
    interest_green_energy_cost=(float(gec) * float(gei))
    interest_battery_cost=(float(bc) * float(bi))
    interest_smarthome_cost=(float(sc) * float(si))
    interest_vehicle_cost=(float(vc) * float(vi))
    interest_bm_cost=(float(bmc) * float(bmi))
    interest_maintenance_cost=(float(mc) * float(mi))
    insurance_maintenance_cost =(float(ic) * float(ii))

    #Summatory
    total_energy_cost = ptc + etc  # It's the same for real and interest cost
    total_real_cost=float(gec)+float(bc)+float(sc)+float(vc)+float(bmc)+float(mc)+float(ic)
    total_interest_cost = interest_green_energy_cost + interest_battery_cost + interest_smarthome_cost \
                        + interest_vehicle_cost + interest_bm_cost + interest_maintenance_cost + insurance_maintenance_cost

    #knapsack algorithm with customer data
    solution_customer=knapsack_algorithm(interest_green_energy_cost,interest_battery_cost,interest_smarthome_cost,interest_vehicle_cost,
                                         interest_bm_cost,interest_maintenance_cost,insurance_maintenance_cost,total_interest_cost)
    solutionToCSV(customer_form_id,total_real_cost,total_interest_cost,total_energy_cost,solution_customer, customer_or_cluster,
                  tariff, power, consumption, customer_type, funding_duration, green_energy_interest, battery_interest,
                  smarthome_interest, vehicle_interest, bm_interest, maintenance_interest, insurance_interest)

    # knapsack algorithm with customer data
    solution_customer = knapsack_algorithm(interest_green_energy_cost, interest_battery_cost, interest_smarthome_cost,interest_vehicle_cost,
                                           interest_bm_cost, interest_maintenance_cost,insurance_maintenance_cost, total_interest_cost)
    solutionToCSV(customer_form_id,total_real_cost,total_interest_cost,total_energy_cost,solution_customer, customer_or_cluster,
                  tariff, power, consumption, customer_type, funding_duration, green_energy_interest, battery_interest,
                  smarthome_interest, vehicle_interest, bm_interest, maintenance_interest, insurance_interest)

    return 1
```

Figure 13: Cost function

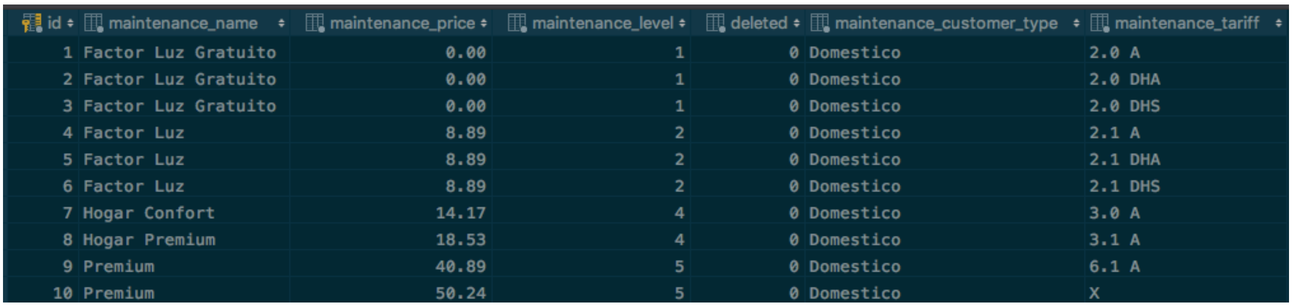
To calculate that this function uses the numeric points of interest of a customer translated in percentages of interest that move between a minimum and a maximum depending on which point the customer choose.

The above script takes the value of interest of each customer and translate it to a percentage using random value generator between the “min_interest” and the “max_interest” that can be seen in the table below. This allows to give some uncertainty to the algorithm because despite a customer can do the survey two times and answer the same, the interest understanding for each subservice will change. Once the interest is computed the algorithm takes the real cost of each subservice, for example for maintenance cost the script is the next one:



interest_value	min_interest	max_interest	deleted
1	0.00	0.19	0
2	0.20	0.39	0
3	0.40	0.59	0
4	0.60	0.79	0
5	0.80	0.99	0

Table 6: Interest table (database)

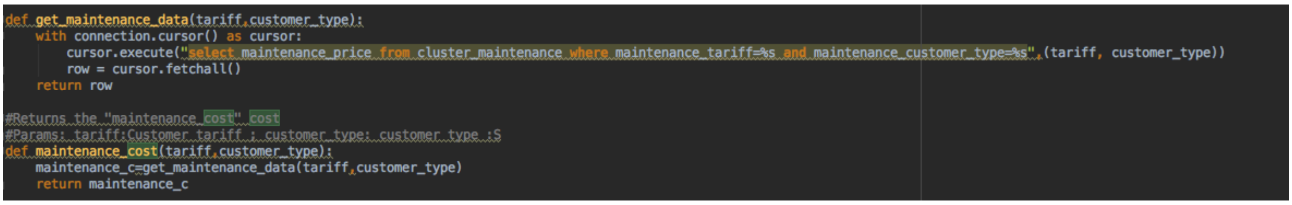


id	maintenance_name	maintenance_price	maintenance_level	deleted	maintenance_customer_type	maintenance_tariff
1	Factor Luz Gratuito	0.00	1	0	Domestico	2.0 A
2	Factor Luz Gratuito	0.00	1	0	Domestico	2.0 DHA
3	Factor Luz Gratuito	0.00	1	0	Domestico	2.0 DHS
4	Factor Luz	8.89	2	0	Domestico	2.1 A
5	Factor Luz	8.89	2	0	Domestico	2.1 DHA
6	Factor Luz	8.89	2	0	Domestico	2.1 DHS
7	Hogar Confort	14.17	4	0	Domestico	3.0 A
8	Hogar Premium	18.53	4	0	Domestico	3.1 A
9	Premium	40.89	5	0	Domestico	6.1 A
10	Premium	50.24	5	0	Domestico	X

Table 7: Maintenance table (database)

This takes the data from the following table.

The rest of costs are computed as the same way but using different tables. Once these are achieved the idea behind the knapsack capacity is to sum all these costs, but if we do it like that the cost achieved is the real cost of a service with all the subservices. Then we must limit this someway, because the customer availability to pay is related to interest.



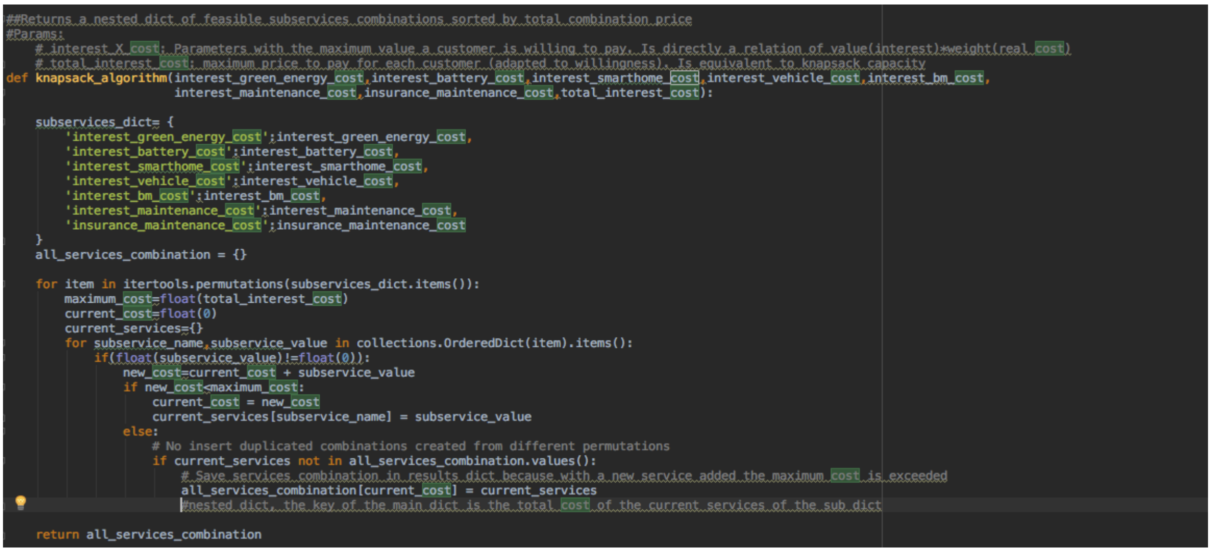
```
def get_maintenance_data(tariff,customer_type):
    with connection.cursor() as cursor:
        cursor.execute("select maintenance_price from cluster_maintenance where maintenance_tariff=%s and maintenance_customer_type=%s",(tariff, customer_type))
        row = cursor.fetchall()
    return row

#Returns the "maintenance_cost" cost
#Params: tariff:Customer tariff ; customer_type: customer type :S
def maintenance_cost(tariff,customer_type):
    maintenance_c=get_maintenance_data(tariff,customer_type)
    return maintenance_c
```

Figure 15: Subservice cost computing example (maintenance)

To do that the formula is the next one:

Subservice Interest Cost = Subservice Real Cost * Customer Interest

For example, if a subservice cost 100 euros, but the customer is interested in a 20%, the maximum willingness to pay for that will be 20 euros. When this is applied to the whole product roster and sum up, you have that the willingness to pay of a customer is for example 200 euros. This will be the knapsack capacity. Any subservice added that exceed this limit will be discarded.

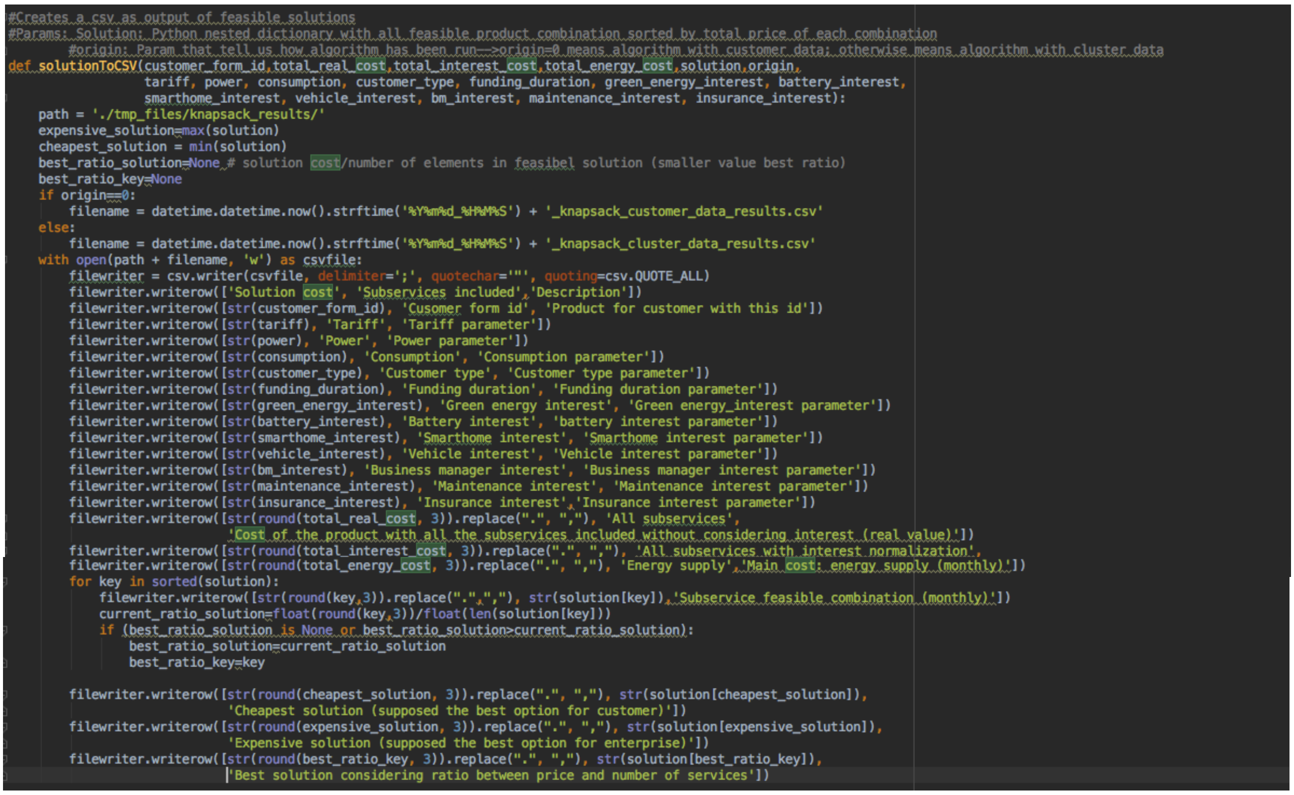


```python
##Returns a nested dict of feasible subservices combinations sorted by total combination price
#Params:
    # interest_X_cost: Parameters with the maximum value a customer is willing to pay. Is directly a relation of value(interest)*weight(real_cost)
    # total_interest_cost: maximum price to pay for each customer (adapted to willingness). Is equivalent to knapsack capacity
def knapsack_algorithm(interest_green_energy_cost,interest_battery_cost,interest_smarthome_cost,interest_vehicle_cost,interest_bm_cost,
                       interest_maintenance_cost,insurance_maintenance_cost,total_interest_cost):

    subservices_dict= {
        'interest_green_energy_cost':interest_green_energy_cost,
        'interest_battery_cost':interest_battery_cost,
        'interest_smarthome_cost':interest_smarthome_cost,
        'interest_vehicle_cost':interest_vehicle_cost,
        'interest_bm_cost':interest_bm_cost,
        'interest_maintenance_cost':interest_maintenance_cost,
        'insurance_maintenance_cost':insurance_maintenance_cost
    }
    all_services_combination = {}

    for item in itertools.permutations(subservices_dict.items()):
        maximum_cost=float(total_interest_cost)
        current_cost=float(0)
        current_services={}
        for subservice_name,subservice_value in collections.OrderedDict(item).items():
            if(float(subservice_value)!=float(0)):
                new_cost=current_cost + subservice_value
                if new_cost<maximum_cost:
                    current_cost = new_cost
                    current_services[subservice_name] = subservice_value
                else:
                    # No insert duplicated combinations created from different permutations
                    if current_services not in all_services_combination.values():
                        # Save services combination in results dict because with a new service added the maximum cost is exceeded
                        all_services_combination[current_cost] = current_services
                        #nested dict, the key of the main dict is the total cost of the current services of the sub dict

    return all_services_combination
```

Figure 16: Knapsack algorithm

Then, one you have iterated over the all set of products, the algorithm has a feasible solution. The “magic” on it is that previously knapsack algorithm has calculated all the product combinatorial, then it will iterate N times to achieve to compute all the feasible solutions.

These solutions are stored in a csv using the next function:



```python
#Creates a csv as output of feasible solutions
#Params: Solution: Python nested dictionary with all feasible product combination sorted by total price of each combination
        #origin: Param that tell us how algorithm has been run-->origin=0 means algorithm with customer data; otherwise means algorithm with cluster data
def solutionToCSV(customer_form_id,total_real_cost,total_interest_cost,total_energy_cost,solution,origin,
                  tariff, power, consumption, customer_type, funding_duration, green_energy_interest, battery_interest,
                  smarthome_interest, vehicle_interest, bm_interest, maintenance_interest, insurance_interest):
    path = './tmp_files/knapsack_results/'
    expensive_solution=max(solution)
    cheapest_solution = min(solution)
    best_ratio_solution=None # solution cost/number of elements in feasibel solution (smaller value best ratio)
    best_ratio_key=None
    if origin==0:
        filename = datetime.datetime.now().strftime('%Y%m%d_%H%M%S') + '_knapsack_customer_data_results.csv'
    else:
        filename = datetime.datetime.now().strftime('%Y%m%d_%H%M%S') + '_knapsack_cluster_data_results.csv'
    with open(path + filename, 'w') as csvfile:
        filewriter = csv.writer(csvfile, delimiter=';', quotechar='"', quoting=csv.QUOTE_ALL)
        filewriter.writerow(['Solution cost', 'Subservices included','Description'])
        filewriter.writerow([str(customer_form_id), 'Cusomer form id', 'Product for customer with this id'])
        filewriter.writerow([str(tariff), 'Tariff', 'Tariff parameter'])
        filewriter.writerow([str(power), 'Power', 'Power parameter'])
        filewriter.writerow([str(consumption), 'Consumption', 'Consumption parameter'])
        filewriter.writerow([str(customer_type), 'Customer type', 'Customer type parameter'])
        filewriter.writerow([str(funding_duration), 'Funding duration', 'Funding duration parameter'])
        filewriter.writerow([str(green_energy_interest), 'Green energy interest', 'Green energy_interest parameter'])
        filewriter.writerow([str(battery_interest), 'Battery interest', 'battery interest parameter'])
        filewriter.writerow([str(smarthome_interest), 'Smarthome interest', 'Smarthome interest parameter'])
        filewriter.writerow([str(vehicle_interest), 'Vehicle interest', 'Vehicle interest parameter'])
        filewriter.writerow([str(bm_interest), 'Business manager interest', 'Business manager interest parameter'])
        filewriter.writerow([str(maintenance_interest), 'Maintenance interest', 'Maintenance interest parameter'])
        filewriter.writerow([str(insurance_interest), 'Insurance interest','Insurance interest parameter'])
        filewriter.writerow([str(round(total_real_cost, 3)).replace(".", ","), 'All subservices',
                             'Cost of the product with all the subservices included without considering interest (real value)'])
        filewriter.writerow([str(round(total_interest_cost, 3)).replace(".", ","), 'All subservices with interest normalization',
        filewriter.writerow([str(round(total_energy_cost, 3)).replace(".", ","), 'Energy supply','Main cost: energy supply (monthly)'])
        for key in sorted(solution):
            filewriter.writerow([str(round(key,3)).replace(".",","), str(solution[key]),'Subservice feasible combination (monthly)'])
            current_ratio_solution=float(round(key,3))/float(len(solution[key]))
            if (best_ratio_solution is None or best_ratio_solution>current_ratio_solution):
                best_ratio_solution=current_ratio_solution
                best_ratio_key=key

        filewriter.writerow([str(round(cheapest_solution, 3)).replace(".", ","), str(solution[cheapest_solution]),
                             'Cheapest solution (supposed the best option for customer)'])
        filewriter.writerow([str(round(expensive_solution, 3)).replace(".", ","), str(solution[expensive_solution]),
                             'Expensive solution (supposed the best option for enterprise)'])
        filewriter.writerow([str(round(best_ratio_key, 3)).replace(".", ","), str(solution[best_ratio_key]),
                             'Best solution considering ratio between price and number of services'])
```

Figure 17: CSV Output generator

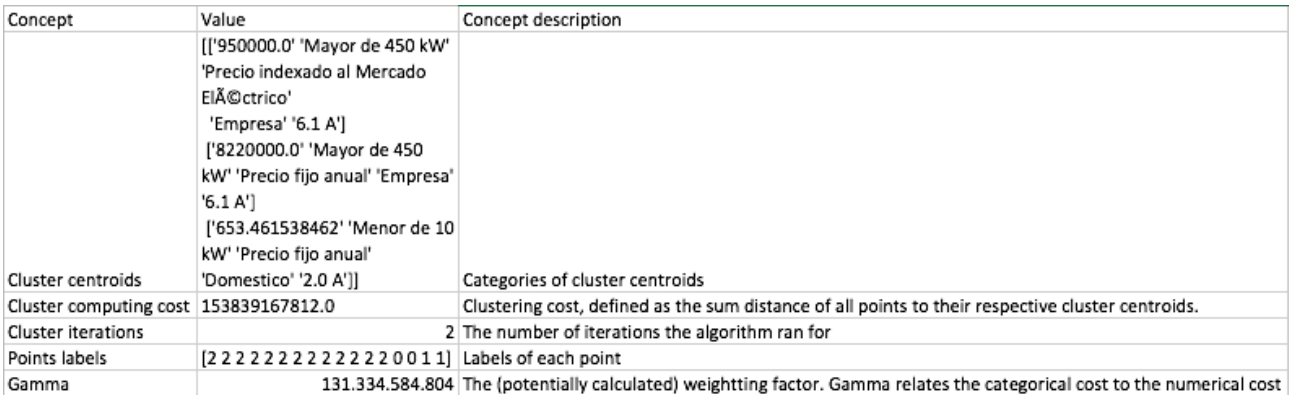
An example of this outputs can be found on the attached documents and this will be analysed on the next section.

6. Testing and result analysis

This section explains how the algorithm has been tested and the results has been analysed.
Every scientific/development project needs after the project is finished a complete validation of what have been done.

In this case the testing phase has been done using retrospective. The output must ensure that a set of restrictions are accomplished.

The output has two different csv that allows to check if the test is successful and helps to understand the results.
In order to check clustering algorithm, the output for one execution is:



Concept	Value	Concept description
Cluster centroids	[['950000.0' 'Mayor de 450 kW' 'Precio indexado al Mercado ElÃ©ctrico' 'Empresa' '6.1 A'] ['8220000.0' 'Mayor de 450 kW' 'Precio fijo anual' 'Empresa' '6.1 A'] ['653.461538462' 'Menor de 10 kW' 'Precio fijo anual' 'Domestico' '2.0 A']]	Categories of cluster centroids
Cluster computing cost	153839167812.0	Clustering cost, defined as the sum distance of all points to their respective cluster centroids.
Cluster iterations	2	The number of iterations the algorithm ran for
Points labels	[2 2 2 2 2 2 2 2 2 2 2 2 2 0 0 1 1]	Labels of each point
Gamma	131.334.584.804	The (potentially calculated) weightting factor. Gamma relates the categorical cost to the numerical cost

Table 8: Output example

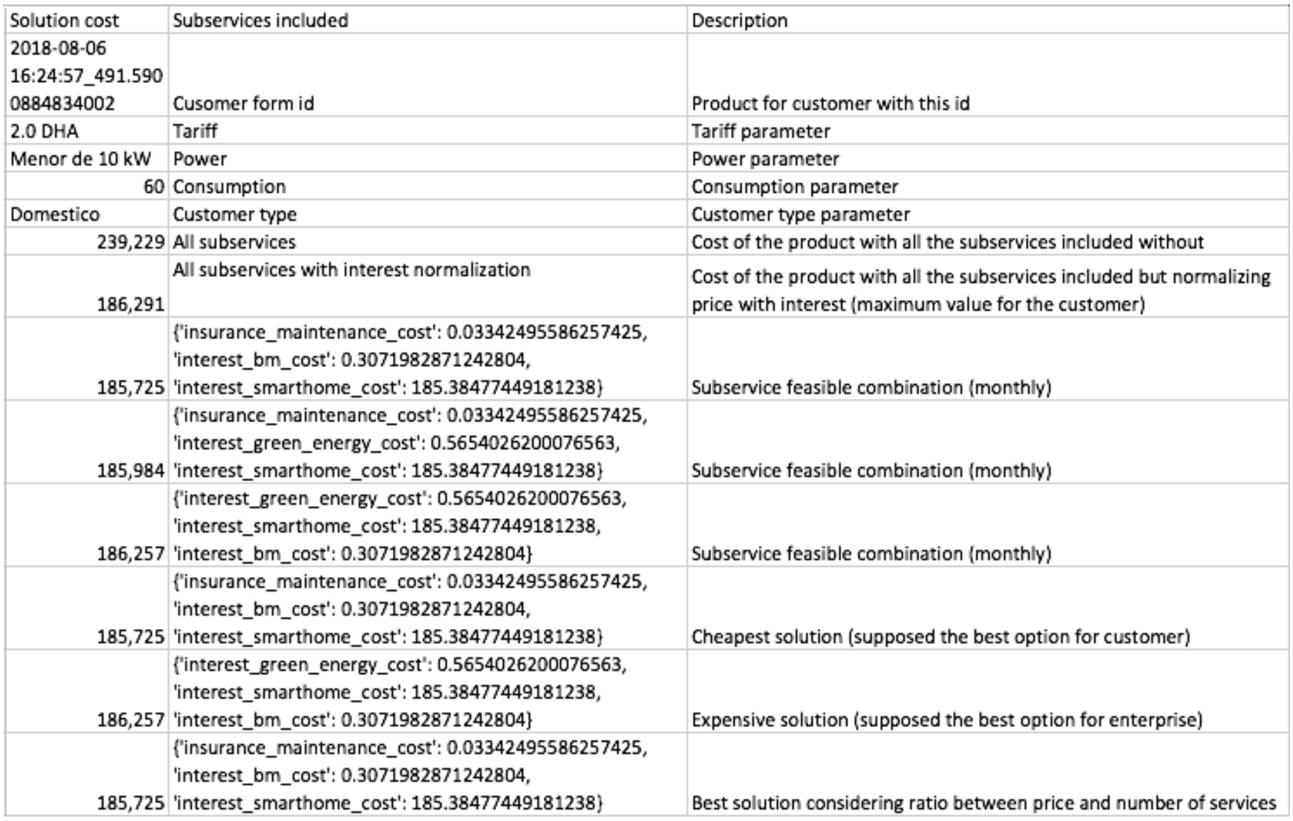
Analysing the centroids is possible to see that the K was three, and that where analysed 17 customers (number of points label). A big part of them, 13, were small customers, this is possible to analyse because the cluster label is 2 for them, and this corresponds for the third cluster centroid that also is in the output csv. This centroid is the one with less hired power, and the smallest tariff.

At this output there is also other key indicators about data set essence. For example, the cluster computing cost indicates how far are the points between them. If this cost is extremely high can be an indicator that maybe a few more clusters can split the data set in more specific clusters. Despite that, this number of clusters also depends of the project nature.

If for example you only have three products, one for big customers, one for small customers, and one customized for VIP customers, then does not make sense create more than three clusters, because some clusters will share products.

Finally, analysing the “Gamma” indicator is possible to see that categorical data has more presence in the data set than numerical. This is because “Gamma” is a big number. It indicates the cost of clustering categorical data versus the cost of clustering numerical data. Other factors can be changed if the customer answers change his nature, for example, if all the customers are close between them the cluster computing cost will be smaller, but Gamma will remain similar because it depends of the variables analysed in the cluster, and this factor is determined by the researchers and developers applying business rules.

Another csv is the knapsack results. This shows the product generation results. In the next tables is possible to watch how it looks for household and enterprise customers.



Solution cost	Subservices included	Description
2018-08-06 16:24:57_491.590 0884834002	Cusomer form id	Product for customer with this id
2.0 DHA	Tariff	Tariff parameter
Menor de 10 kW	Power	Power parameter
60	Consumption	Consumption parameter
Domestico	Customer type	Customer type parameter
239,229	All subservices	Cost of the product with all the subservices included without
186,291	All subservices with interest normalization	Cost of the product with all the subservices included but normalizing price with interest (maximum value for the customer)
185,725	{'insurance_maintenance_cost': 0.03342495586257425, 'interest_bm_cost': 0.3071982871242804, 'interest_smarthome_cost': 185.38477449181238}	Subservice feasible combination (monthly)
185,984	{'insurance_maintenance_cost': 0.03342495586257425, 'interest_green_energy_cost': 0.5654026200076563, 'interest_smarthome_cost': 185.38477449181238}	Subservice feasible combination (monthly)
186,257	{'interest_green_energy_cost': 0.5654026200076563, 'interest_smarthome_cost': 185.38477449181238, 'interest_bm_cost': 0.3071982871242804}	Subservice feasible combination (monthly)
185,725	{'insurance_maintenance_cost': 0.03342495586257425, 'interest_bm_cost': 0.3071982871242804, 'interest_smarthome_cost': 185.38477449181238}	Cheapest solution (supposed the best option for customer)
186,257	{'interest_green_energy_cost': 0.5654026200076563, 'interest_smarthome_cost': 185.38477449181238, 'interest_bm_cost': 0.3071982871242804}	Expensive solution (supposed the best option for enterprise)
185,725	{'insurance_maintenance_cost': 0.03342495586257425, 'interest_bm_cost': 0.3071982871242804, 'interest_smarthome_cost': 185.38477449181238}	Best solution considering ratio between price and number of services

Table 9: Household Customer knapsack output

Firstly, this file includes the cluster variables that has been analysed in the previous csv. It means that this file will be generated for each execution, and each execution refers to a customer, then it is important to know to which customer it refers.

To achieve that Customer id, is included. Moreover, tariff, power, consumption, and customer type are also included. The goal to this is that the engineer that receives this output can make a fast validation that the results, and the expected cost of the product is real and feasible for a customer with these characteristics.

The consultants that has helped in this project require this as a need, and the main reason is because engineers need to do a check of what will be offered to the customer. This is related with the non-complete automatization of the product generation.

Afterwards, the next rows contain the customer responses about interest and the total real cost. If the interest is high enough for all of the subservices, the "total_interest_cost" that is the cost that represents the knapsack capacity will be almost identical to "total_real_cost". It is simply the sum of all subservices cost.

Another cost in this output that has not been told in the complete project is "total_energy_cost". The main business of marketers is energy supply, then all the other subservices are extras. Whether finally the customer is not interested in the product offer that the algorithm has generated for them, and the seller or engineer cannot adjust the offer for them, can be interesting to sell almost the main product of the company. A customer in our company is a customer that competence do not have.

Lastly, the last three lines includes the real output of the product generation execution. They are the cheapest solution, it is supposed to be the best option for customer. The expensive solution, it is supposed the best option for enterprise. And finally, the best solution considering ratio between price and number of services, a lot of customers evaluates more the number of services that their "knapsack" includes than the quality of them. So, can be interesting to offer a greater number of subservices with cheaper prices, and not two or three subservices that are high quality.

This product generation algorithm needs two components to be tested. One is the customer that has answered the survey and asks if at least one of the three product lines generated accomplish to satisfy their needs. The other one is the consultants that check if this product fit the business requirements they indicate.

Obviously, the first one is hard to achieve, because the survey was online, and it does not ask for contact data due to the investigation nature of the project, the change on GDPR, and the already few people that has answered them once the questions were limited and the poll do not implicate further steps.

Then, to simulate that, the researcher asks a few acquaintances of the work. They did the survey, and then the researches have provided what the algorithm generated for them.
Globally the results were good because the price do not increase more than what they really are able to pay, and they have two different options of product, the cheapest one, and the best ratio. Anyone consider paying for the expensive product, but this was expected.

Is interesting that the maximum availability to pay was fair enough because the survey does not ask in any moment for this willingness. Then the technique to use to estimate that seems to be good, at least for this first approach.

Some of them regrets about that some product does not include some service that they really interest. It is due the combinatorial nature of the product generation process. If you are interested in one product that cost 50 euros, and your limit is 100 euros, and the rest of services cost 21 euros. The cheapest solution will be 4 subservices of this last price (82 euros) and not another combination that includes the product that you are really interested, an example is that this product plus two of 21 euros, sum 92 euros, then the other one is cheaper, do not break the limit, and is the best ratio solution.
Consultants also find this flaw.

The consultants checked if the products fit the business rules. To do that they consider a few things:

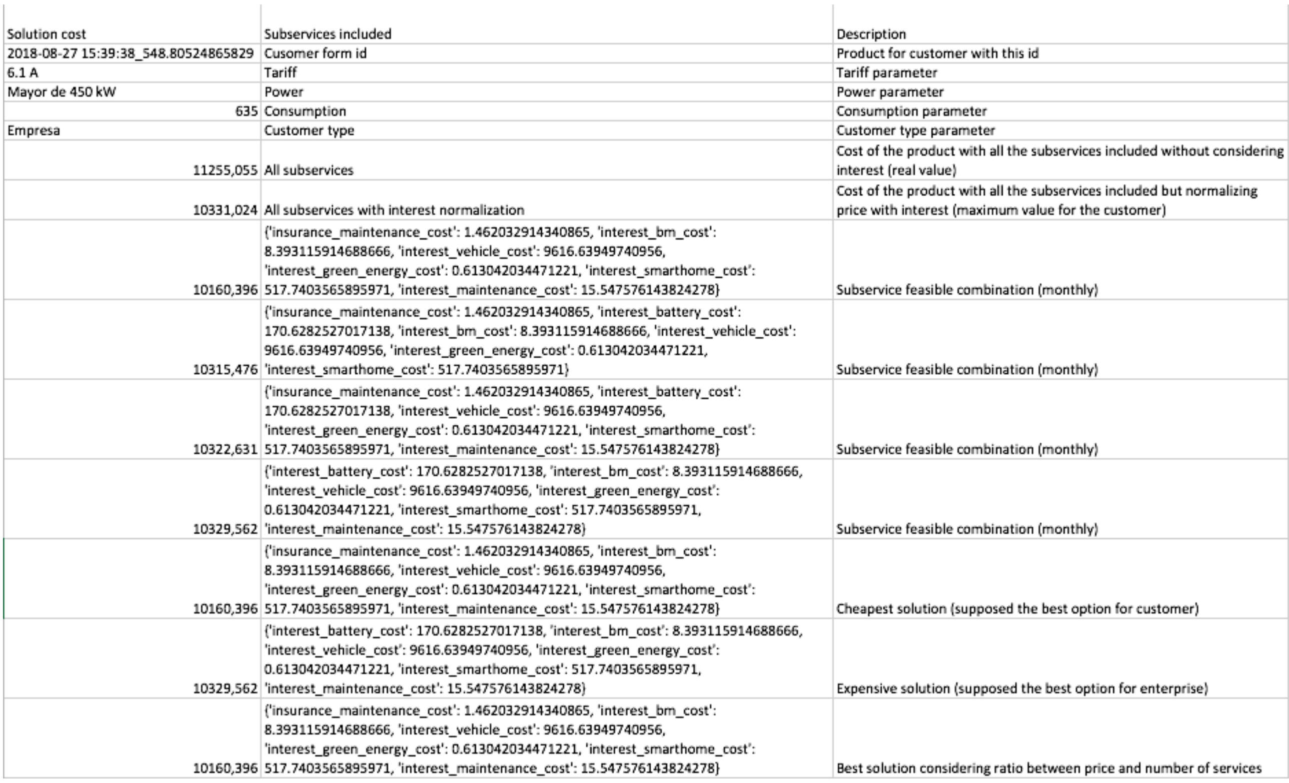
- The total real cost
- The solution costs
- The customers answer

Consultants have years of experience in this sector and then, they know when a product has an overprice for what a customer can find on competence. To check that they use total real cost, they know that this price will be high, and probably any customer will pay for it, but using this is possible to know if the product generated for a customer corresponds to his category.

An example of it is that a total real cost for an enterprise can be 4 or 5 time bigger than a household customer, so when the tariff, the customer type, the hired power and the consumption is
for small customer certain ranges of prices does not make sense and this is what consultants validate here.

The insurance of a house is much cheaper that a company, and moreover when a company has different levels of insurance available.

This first validation was successfully tested as can be seen on the next screenshots (attached in files).



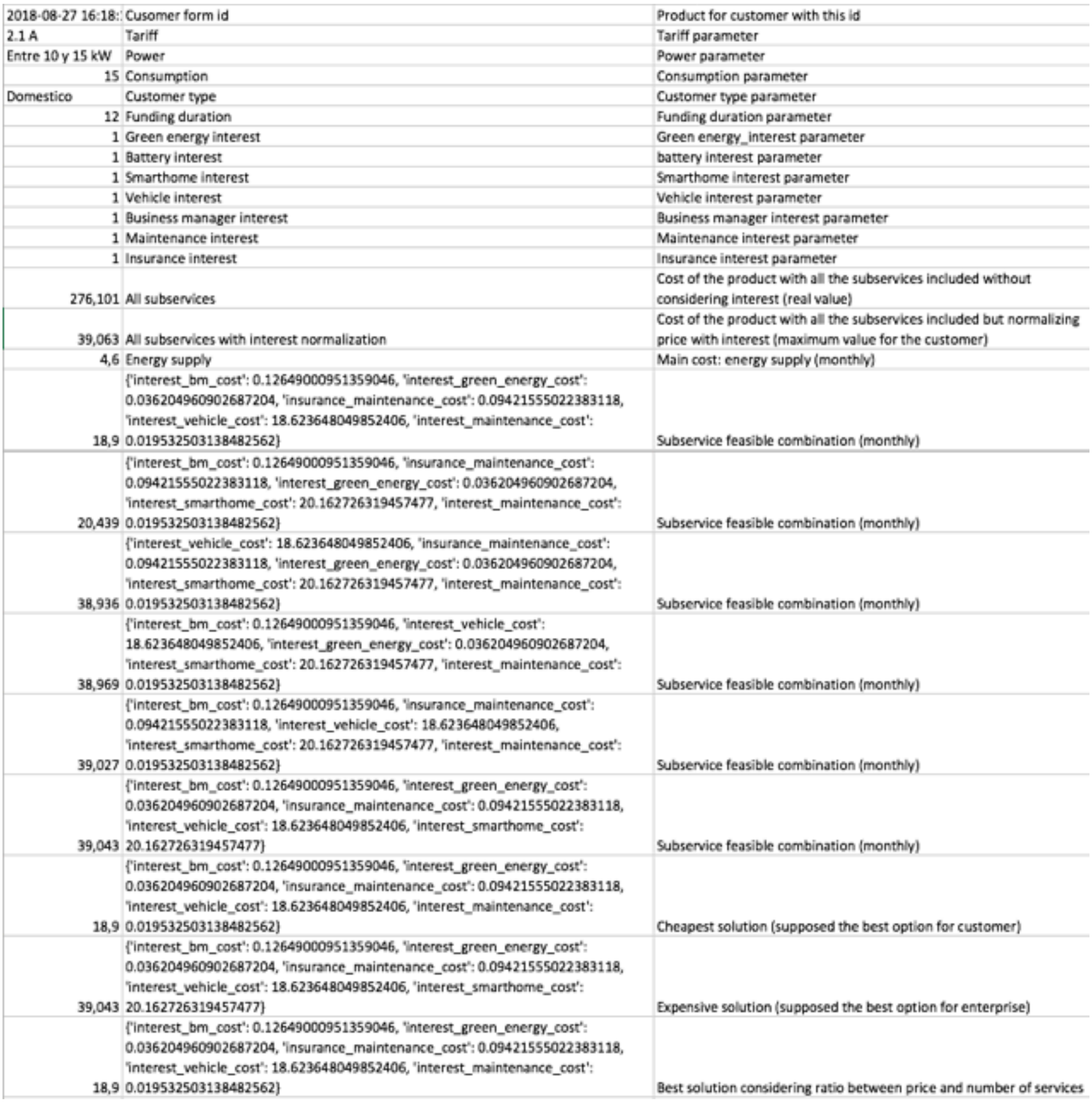
Solution cost	Subservices included	Description
2018-08-27 15:39:38_548.80524865829	Cusomer form id	Product for customer with this id
6.1 A	Tariff	Tariff parameter
Mayor de 450 kW	Power	Power parameter
635	Consumption	Consumption parameter
Empresa	Customer type	Customer type parameter
11255,055	All subservices	Cost of the product with all the subservices included without considering interest (real value)
10331,024	All subservices with interest normalization	Cost of the product with all the subservices included but normalizing price with interest (maximum value for the customer)
10160,396	{'insurance_maintenance_cost': 1.462032914340865, 'interest_bm_cost': 8.393115914688666, 'interest_vehicle_cost': 9616.63949740956, 'interest_green_energy_cost': 0.613042034471221, 'interest_smarthome_cost': 517.7403565895971, 'interest_maintenance_cost': 15.547576143824278}	Subservice feasible combination (monthly)
10315,476	{'insurance_maintenance_cost': 1.462032914340865, 'interest_battery_cost': 170.6282527017138, 'interest_bm_cost': 8.393115914688666, 'interest_vehicle_cost': 9616.63949740956, 'interest_green_energy_cost': 0.613042034471221, 'interest_smarthome_cost': 517.7403565895971}	Subservice feasible combination (monthly)
10322,631	{'insurance_maintenance_cost': 1.462032914340865, 'interest_battery_cost': 170.6282527017138, 'interest_vehicle_cost': 9616.63949740956, 'interest_green_energy_cost': 0.613042034471221, 'interest_smarthome_cost': 517.7403565895971, 'interest_maintenance_cost': 15.547576143824278}	Subservice feasible combination (monthly)
10329,562	{'interest_battery_cost': 170.6282527017138, 'interest_bm_cost': 8.393115914688666, 'interest_vehicle_cost': 9616.63949740956, 'interest_green_energy_cost': 0.613042034471221, 'interest_smarthome_cost': 517.7403565895971, 'interest_maintenance_cost': 15.547576143824278}	Subservice feasible combination (monthly)
10160,396	{'insurance_maintenance_cost': 1.462032914340865, 'interest_bm_cost': 8.393115914688666, 'interest_vehicle_cost': 9616.63949740956, 'interest_green_energy_cost': 0.613042034471221, 'interest_smarthome_cost': 517.7403565895971, 'interest_maintenance_cost': 15.547576143824278}	Cheapest solution (supposed the best option for customer)
10329,562	{'interest_battery_cost': 170.6282527017138, 'interest_bm_cost': 8.393115914688666, 'interest_vehicle_cost': 9616.63949740956, 'interest_green_energy_cost': 0.613042034471221, 'interest_smarthome_cost': 517.7403565895971, 'interest_maintenance_cost': 15.547576143824278}	Expensive solution (supposed the best option for enterprise)
10160,396	{'insurance_maintenance_cost': 1.462032914340865, 'interest_bm_cost': 8.393115914688666, 'interest_vehicle_cost': 9616.63949740956, 'interest_green_energy_cost': 0.613042034471221, 'interest_smarthome_cost': 517.7403565895971, 'interest_maintenance_cost': 15.547576143824278}	Best solution considering ratio between price and number of services

Table 10: Enterprise Customer knapsack output

On the other hand, the solution cost and the customer answers are related for consultants. If the interest is low for the big part of subservices the solution cost should be low, because if it is not will be hard to sell it to the customers.

They found that the relation between interest and solution cost is directly dependent (images below) and this generate a good understanding in the algorithm of what a customer wants. So, they approve this. This test checks if the total price of the product including subservice interest, is higher for customers with the maximum interest than for those that answer the minimum one.

The test itself was to answer two times the survey with the same customer data but changing the interest answers. For one survey the interest was one for all subservices, for the other one this interest was five.

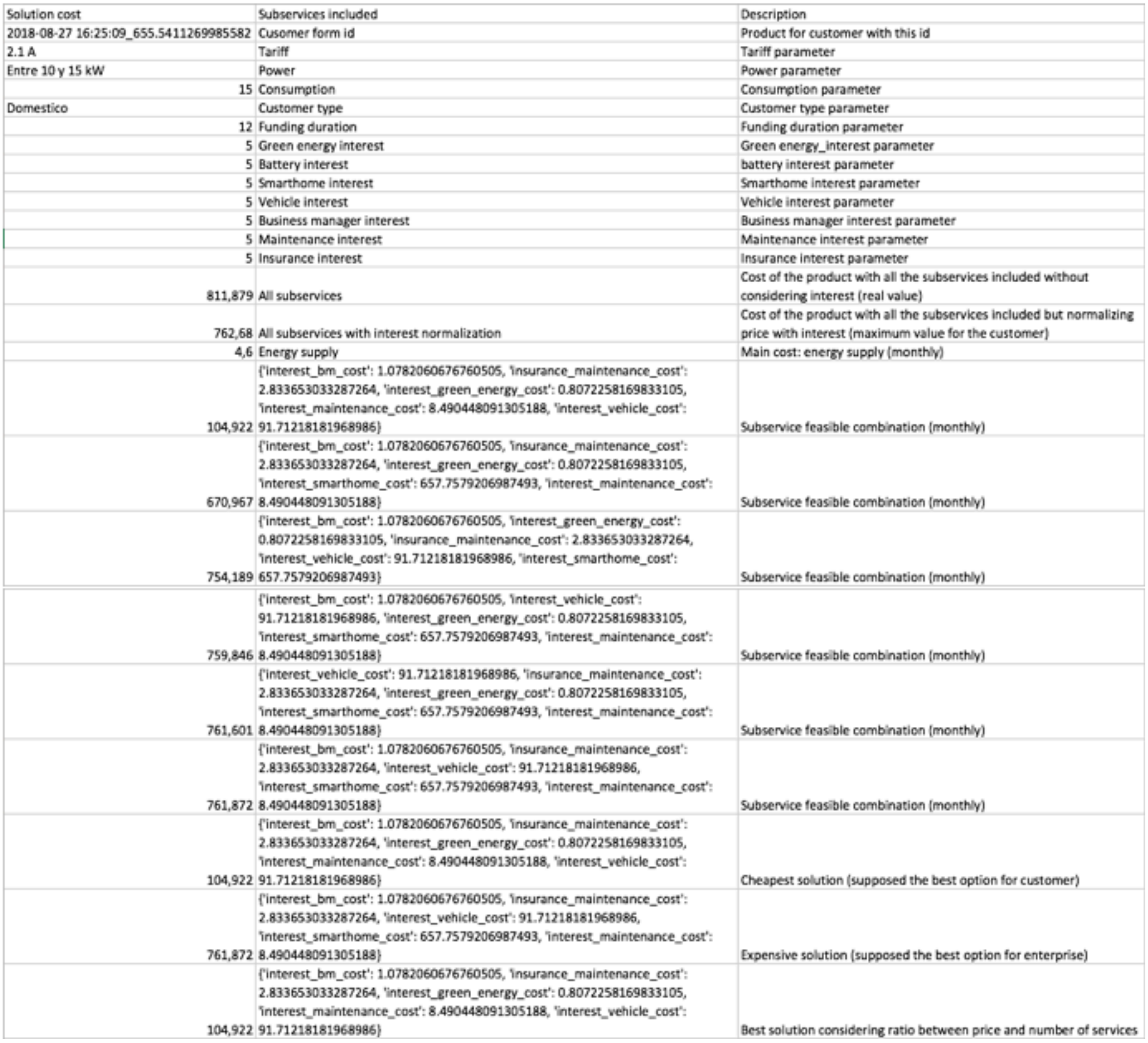


2018-08-27 16:18::	Cusomer form id	Product for customer with this id
2.1 A	Tariff	Tariff parameter
Entre 10 y 15 kW	Power	Power parameter
15	Consumption	Consumption parameter
Domestico	Customer type	Customer type parameter
12	Funding duration	Funding duration parameter
1	Green energy interest	Green energy_interest parameter
1	Battery interest	battery interest parameter
1	Smarthome interest	Smarthome interest parameter
1	Vehicle interest	Vehicle interest parameter
1	Business manager interest	Business manager interest parameter
1	Maintenance interest	Maintenance interest parameter
1	Insurance interest	Insurance interest parameter
276,101	All subservices	Cost of the product with all the subservices included without considering interest (real value)
39,063	All subservices with interest normalization	Cost of the product with all the subservices included but normalizing price with interest (maximum value for the customer)
4,6	Energy supply	Main cost: energy supply (monthly)
18,9	{'interest_bm_cost': 0.12649000951359046, 'interest_green_energy_cost': 0.036204960902687204, 'insurance_maintenance_cost': 0.09421555022383118, 'interest_vehicle_cost': 18.623648049852406, 'interest_maintenance_cost': 0.019532503138482562}	Subservice feasible combination (monthly)
20,439	{'interest_bm_cost': 0.12649000951359046, 'insurance_maintenance_cost': 0.09421555022383118, 'interest_green_energy_cost': 0.036204960902687204, 'interest_smarthome_cost': 20.162726319457477, 'interest_maintenance_cost': 0.019532503138482562}	Subservice feasible combination (monthly)
38,936	{'interest_vehicle_cost': 18.623648049852406, 'insurance_maintenance_cost': 0.09421555022383118, 'interest_green_energy_cost': 0.036204960902687204, 'interest_smarthome_cost': 20.162726319457477, 'interest_maintenance_cost': 0.019532503138482562}	Subservice feasible combination (monthly)
38,969	{'interest_bm_cost': 0.12649000951359046, 'interest_vehicle_cost': 18.623648049852406, 'interest_green_energy_cost': 0.036204960902687204, 'interest_smarthome_cost': 20.162726319457477, 'interest_maintenance_cost': 0.019532503138482562}	Subservice feasible combination (monthly)
39,027	{'interest_bm_cost': 0.12649000951359046, 'insurance_maintenance_cost': 0.09421555022383118, 'interest_vehicle_cost': 18.623648049852406, 'interest_smarthome_cost': 20.162726319457477, 'interest_maintenance_cost': 0.019532503138482562}	Subservice feasible combination (monthly)
39,043	{'interest_bm_cost': 0.12649000951359046, 'interest_green_energy_cost': 0.036204960902687204, 'insurance_maintenance_cost': 0.09421555022383118, 'interest_vehicle_cost': 18.623648049852406, 'interest_smarthome_cost': 20.162726319457477}	Subservice feasible combination (monthly)
18,9	{'interest_bm_cost': 0.12649000951359046, 'interest_green_energy_cost': 0.036204960902687204, 'insurance_maintenance_cost': 0.09421555022383118, 'interest_vehicle_cost': 18.623648049852406, 'interest_maintenance_cost': 0.019532503138482562}	Cheapest solution (supposed the best option for customer)
39,043	{'interest_bm_cost': 0.12649000951359046, 'interest_green_energy_cost': 0.036204960902687204, 'insurance_maintenance_cost': 0.09421555022383118, 'interest_vehicle_cost': 18.623648049852406, 'interest_smarthome_cost': 20.162726319457477}	Expensive solution (supposed the best option for enterprise)
18,9	{'interest_bm_cost': 0.12649000951359046, 'interest_green_energy_cost': 0.036204960902687204, 'insurance_maintenance_cost': 0.09421555022383118, 'interest_vehicle_cost': 18.623648049852406, 'interest_maintenance_cost': 0.019532503138482562}	Best solution considering ratio between price and number of services

Table 11: Knapsack output with interest one in subservices

An objection done by the consultants was the same done by the customers, they find a lack of top interested products in some solutions for customers.

As explained above this is a common point of both analyses, then this improvement will be considered at the next section.



Solution cost	Subservices included	Description
2018-08-27 16:25:09_655.5411269985582	Cusomer form id	Product for customer with this id
2.1 A	Tariff	Tariff parameter
Entre 10 y 15 kW	Power	Power parameter
15	Consumption	Consumption parameter
Domestico	Customer type	Customer type parameter
12	Funding duration	Funding duration parameter
5	Green energy interest	Green energy_interest parameter
5	Battery interest	battery interest parameter
5	Smarthome interest	Smarthome interest parameter
5	Vehicle interest	Vehicle interest parameter
5	Business manager interest	Business manager interest parameter
5	Maintenance interest	Maintenance interest parameter
5	Insurance interest	Insurance interest parameter
811,879	All subservices	Cost of the product with all the subservices included without considering interest (real value)
762,68	All subservices with interest normalization	Cost of the product with all the subservices included but normalizing price with interest (maximum value for the customer)
4,6	Energy supply	Main cost: energy supply (monthly)
104,922	{'interest_bm_cost': 1.0782060676760505, 'insurance_maintenance_cost': 2.833653033287264, 'interest_green_energy_cost': 0.8072258169833105, 'interest_maintenance_cost': 8.490448091305188, 'interest_vehicle_cost': 91.71218181968986}	Subservice feasible combination (monthly)
670,967	{'interest_bm_cost': 1.0782060676760505, 'insurance_maintenance_cost': 2.833653033287264, 'interest_green_energy_cost': 0.8072258169833105, 'interest_smarthome_cost': 657.7579206987493, 'interest_maintenance_cost': 8.490448091305188}	Subservice feasible combination (monthly)
754,189	{'interest_bm_cost': 1.0782060676760505, 'interest_green_energy_cost': 0.8072258169833105, 'insurance_maintenance_cost': 2.833653033287264, 'interest_vehicle_cost': 91.71218181968986, 'interest_smarthome_cost': 657.7579206987493}	Subservice feasible combination (monthly)
759,846	{'interest_bm_cost': 1.0782060676760505, 'interest_vehicle_cost': 91.71218181968986, 'interest_green_energy_cost': 0.8072258169833105, 'interest_smarthome_cost': 657.7579206987493, 'interest_maintenance_cost': 8.490448091305188}	Subservice feasible combination (monthly)
761,601	{'interest_vehicle_cost': 91.71218181968986, 'insurance_maintenance_cost': 2.833653033287264, 'interest_green_energy_cost': 0.8072258169833105, 'interest_smarthome_cost': 657.7579206987493, 'interest_maintenance_cost': 8.490448091305188}	Subservice feasible combination (monthly)
761,872	{'interest_bm_cost': 1.0782060676760505, 'insurance_maintenance_cost': 2.833653033287264, 'interest_vehicle_cost': 91.71218181968986, 'interest_smarthome_cost': 657.7579206987493, 'interest_maintenance_cost': 8.490448091305188}	Subservice feasible combination (monthly)
104,922	{'interest_bm_cost': 1.0782060676760505, 'insurance_maintenance_cost': 2.833653033287264, 'interest_green_energy_cost': 0.8072258169833105, 'interest_maintenance_cost': 8.490448091305188, 'interest_vehicle_cost': 91.71218181968986}	Cheapest solution (supposed the best option for customer)
761,872	{'interest_bm_cost': 1.0782060676760505, 'insurance_maintenance_cost': 2.833653033287264, 'interest_vehicle_cost': 91.71218181968986, 'interest_smarthome_cost': 657.7579206987493, 'interest_maintenance_cost': 8.490448091305188}	Expensive solution (supposed the best option for enterprise)
104,922	{'interest_bm_cost': 1.0782060676760505, 'insurance_maintenance_cost': 2.833653033287264, 'interest_green_energy_cost': 0.8072258169833105, 'interest_maintenance_cost': 8.490448091305188, 'interest_vehicle_cost': 91.71218181968986}	Best solution considering ratio between price and number of services

Table 12: Knapsack output with interest five in subservices

7. Future lines of work

Eventually, after the analysis of the algorithm and the results, the different profiles that has helped in this project development, find some improvements to some part of the projects.

Respecting data gathering, the improvements, and consequently future lines of work are:

1. Include willingness to pay in survey:

   Adding this to the survey will allow to check if the maximum payment expectation computed by the algorithm corresponds to the customer willingness.

This comparison will provide further data to improve the section of the algorithm that estimates the maximum willingness of a customer, and then achieve better results in the product generation.

It is important to have a correct estimation process because some customers will not provide this data, and despite they provide a range in euros, this can be moved depending on what the company offers them and the interest of these customers respect this subservices. Another pro is that later in production this question can be eliminated and then the survey will be smaller, and this is a very important point to take in account, because people do not want to spend a lot of time providing data and answering polls.

2. Include contact data to test the whole cycle:

   Has been difficult to analyse whether the answers provided by the algorithm were good enough or not to the customers. Finally, this could be saved using acquaintance help but is interesting to add feedback of people that has not been involved in the project development.

Respect clustering and product generation algorithm:

1. Introduce priority in subservices:

   Consultants and team mates find the same flaw, sometimes the product suggestion includes a lot of subservices, but not the one that attract more the customer. This is an improvement that future developments of this project should consider. For example, if a subservice has more than 75% of the interest it is a must for the customer, and then the product generation iteration should start with it, and ergo the solution should include it.

2. Include templates product:

   Time is gold, and using this premise, could be a good idea to reduce product generation time. To achieve that is possible to store all the products that has been generated and associate to cluster. Then, using the most sold products is possible to find the subservices that interest more in a specific customer profile and then use one of this as a template for each cluster, and keep redoing this analysis because in time line this can change.

Once a customer answer a survey, the cluster can be computed and the first suggestion can be the template, if it fits them, is possible to automatically sell this product, and if it is not, then asks the customer for contact data and that a deeper study will be done. Using this, the contact data can be subtracted of the survey, and maybe will be easier to ask for it uses the need to contact with them to explain the product generation study that the algorithm has been done.

8. Managerial Insights

This Computational Marketing project develops a methodology to improve the product generation into the energy marketers companies. This area is performed by the call centre department, who retrieves and gathers data about the potential lead customer, and the technical engineering department, who analyse which customers deserve a deeper study and consequently a deeper study about which product can fit they.

Firstly, is important to understand that this process is fully manual, and depend on the human factor, so the analysis component is limited. This means that not all customers are analysed, and the engineering department apply before nothing big restrictions to clean the pool of customers who receive the study.

Another important question is that call centre do the interview with the customer and this interview can communicate some questions in different ways depending on the agent, the humour sense and others. This affect in big part in the way that is communicated the questions to the customer and it can affect the answers, if this answers is not provided by the customer itself the reliability of the algorithm will decrease.

Both points are solved by the automatization of this process. All customers receive exactly the same questions performed in the same way, no human intercede on the answer process, then the reliability of the poll is bigger than using other communication channels. Once this questions are received, the clustering performs a customer analysis that allows to create a customized product for every customer not for a small subset of them, this approach allows to have more potential leads, so in fewer time there is more customers to catch.

In average when a customer is interested in a company the phone call has a duration of 7 minutes. They ask some questions and receive some other questions, with these a lead is created and then, a person in the technical department take a look of the leads that pass some conditions, for example, the most typical is tariff 3.0A. Once this is accomplished, they look carefully for this customer searching bills, and other information provided by the suppliers in the market (Sips files). If all parameters are okay then using excel do some calculations, and with these and their experience they manipulate some product template until achieve a product that can fit the customer. Moreover, this last part is done speaking by phone to the customer and receiving direct input from the opinion about the product.

All this steps except reaching the customer with an engineer that can do commercial tasks an adjust the product can be done using the developed tool. Then, in fewer time, using less resources the number of customers analysed are bigger, and both departments can focus they work into some other tasks that have more value to the enterprise and has hardly automatization process.
Additionally, a lot of human factor is decreased, then human errors and human manipulation is practically erased from the process, creating a clean funnel.

9. Conclusions

Computational Marketing draws on several disciplines within computer science, and economic theory. The main components are:

1. Information retrieval: To analyse data, first you need this data.
    a. Machine learning: Artificial intelligence branch that creates ways to recognize and react to complex patterns.
    b. Optimization: On one hand, is important to maximize the accessibility of a user into a website, but also machine learning uses heuristics that involves optimization of a goal function.
    c. Microeconomics: studies the behaviour of individuals and firms in making decisions regarding the allocation of resources. In this case, both, enterprise and customer should take decisions.

Then, the project can be considered successful in terms of Computational Marketing scope if these main components have been managed.
As can be seen previously the four pillars were treated in the project scope using surveys, algorithms related to machine learning and optimization, and finally involving the customer and the engineers in the product generation process.
In terms of goal accomplishment, the project can be considered successful too.

The partial goals were:

1. Study the automatic generation product problem and the respective theorical and practical techniques developed by other researchers.
2. Process definition to solve the problem.
3. Algorithm development:
4. Database development
5. Design and create a set of experiments

The first four points have been fully achieved without a problem. Witness of it are the literature review, the algorithm explanations, and the PostgreSQL database set up using Docker. Despite that, the fifth point needs complete interaction with consultants and the customers that has helped in the development. Due to this reason, the researcher does not consider this goal as achieved and let it as future work development for new iterations in the project development.

Even thought, is feasible to consider the main goal - develop a machine learning algorithm in Python to help energy sector enterprises in the classification of their customers and in the automatic product generation for these - as achieved.

As seen in the managerial insights section the contribution of this investigation is to add value in the product generation chain in the marketers of energy sector.

Nowadays the process implies different departments and a lot of manual tasks that creates a great dependence on human knowledge. This also creates space for human error and incompatibility to study all the particular cases. Then the customer pool is not profit at all. Business is then losing opportunities, and maybe not offering what the customer really desire.

The project develops a system and a tool to perform better this process and all using and automatic method, taking profit of all potential leads and decreasing the total working time of the different departments. This creates an optimisation in the process and consequently in the company. The same employees can touch more customers, so the enterprise has the possibility to earn more customers. On the other hand, the workers performs their tasks with the support of a software that helps to reduce error, making easy the decision making process and improving their own results and in fact the company results.

Finally, using these previous reflections as starting point is possible to extract some lessons to apply in future projects:

1. The project planning is important. Despite that, the high number of tasks to develop and the incompatibility with some personal aspects, has been delaying the project too long.

2. The development can use customer data, and their feedback but it is important to keep the traceability in the whole cycle to ensure a correct testing phase.

This lessons glimpse that the original planning has not been followed due to different reasons, despite that is important to remark the great adaptability shown by the university and the project developers. Furthermore, this adaptability has been combined with a good methodology of development and the result is the expected one.

Finally, there are some interesting future lines of work that keeps some doors open for future developments and project improvements to make easier their accommodation to any enterprise.

10. Glossary

Algorithm: A set of step-by-step instructions. Computer algorithms can be simple (if it's 3 p.m., send a reminder) or complex (identify pedestrians).

Artificial Intelligence: Computer programs designed to solve difficult problems which humans (and animals) routinely solve. In a nutshell, is to enable computers to think and learn-by-itself through feeding of good data. The goal of AI is to develop programs which can solve such problems independently, although the patterns for solving these problems differ significantly from the way they are solved by humans.

Data Science: It is a study that unifies statistics, data modelling & visualization, and analysis to extract information, classify data etc.

Classifiers: Algorithms (like., KNN, SVM) used for data classification machine learning.

Machine Learning: A subset of AI in which computer programs and algorithms can be designed to “learn” how to complete a specified task, with increasing efficiency and effectiveness as it develops. Such programs can use past performance data to predict and improve future performance.

Reinforcement learning algorithms: A type of machine learning in which machines are “taught” to achieve their target function through a process of experimentation and reward. In reinforcement learning, the machine receives positive reinforcement when its processes produce the desired result, and negative reinforcement when they do not.

Supervised learning algorithms: A type of machine learning in which human input and supervision are an integral part of the machine learning process on an ongoing basis. In supervised learning, there is a clear outcome to the machine's data mining and its target function is to achieve this outcome, nothing more.

Training Data: In machine learning, the training data set is the data given to the machine during the initial "learning" or "training" phase. From this data set, the machine is meant to gain some insight into options for the efficient completion of its assigned task through identifying relationships between the data.

Unsupervised learning algorithms: A type of machine learning in which human input and supervision are extremely limited, or absent altogether, throughout the process. In unsupervised learning, the machine is left to identify patterns and draw its own conclusions from
the data sets it is given.

Data mining: The examination of data sets to discover and mine patterns from that data that can be of further use.

Cluster analysis: A type of unsupervised learning used for exploratory data analysis to find hidden patterns or grouping in data; clusters are modelled with a measure of similarity defined by metrics such as Euclidean or probabilistic distance.

Clustering: Clustering algorithms let machines group data points or items into groups with similar characteristics.

k-means clustering: It is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining.

K-modes extends the k-means paradigm to cluster categorical data by using a simple matching dissimilarity measure for categorical objects, modes instead of means for clusters and a frequency-based method to update modes in the k-means fashion to minimize the clustering cost function of clustering.

Heuristic search techniques: Support that narrows down the search for optimal solutions for a problem by eliminating options that are incorrect.

11. Bibliography

Book:

- Willi Richert, Luis Pedro Coelho - Building Machine Learning Systems with Python - PACKT publishing - Birmingham – 2013
- Fabrizio Romano - Learning Python - PACKT publishing - Birmingham – 2015
- Sunila Gollapudi - Practical machine learning - PACKT publishing - Birmingham – 2016
- John Hearty - Advanced Machine Learning with Python - PACKT publishing - Birmingham – 2016
- Gabriel Cánepa - What you need to know about machine learning - PACKT publishing - Birmingham - 2016

Papers:

- Huang, Z.: Clustering large data sets with mixed numeric and categorical values, Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore, pp. 21-34, 1997.
- Huang, Z.: Extensions to the k-modes algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery 2(3), pp. 283-304, 1998.
- Z. Gu, M. Xi Tang, J.H. Frazer - Capturing aesthetic intention during interactive evolution - Comput Aided December - 2006 (224-237).
- J. Kuang, P.Jiang - Product platform design for a product family based on Kansei engineering - J. Ena December 20 (6) - 2009 (589-607).
- Cao, F., Liang, J, Bai, L.: A new initialization method for categorical data clustering, Expert Systems with Applications 36(7), pp. 10223-10228., 2009.
- Thomas Y. Lee, Eric T. BradLow - Automated Marketing Research Using Online Customer Reviews - Journal of Marketing Research - 2011
- Eric Lutters,Fred J.A.M.van Houten, Alain Bernard, Emmanuel Mermoz, Corné S.L.Schutte - Tools and techniques for product design- CIRP Annals - 2014
- Zhenyu Zhang, Qingjin Peng, Peihua Gu - Improvement of User Involvement in Product Design - Procedia CIRP - 2015
- Salman Nazari-Shirkouhia,Abbas Keramati - Modeling customer satisfaction with new product design using a flexible fuzzy regression-data envelopment analysis algorithm - Applied Mathematical Modelling - 2017
- Karen Holtzblatt, Hugh Beyer - The Challenge of Product Design - Contextual Design – 2017.

12. Annexes

Interview template

1. First of all, can you introduce yourself? (A brief explanation of your studies and your professional career)
2. Do you know about machine learning and artificial intelligence? (what is it, its applications, etc...)
3. Do you know about automatic product generation algorithms? In case of a positive answer, do you know if any energy sector enterprise is trying to develop an algorithm with this purpose? Have you been involved in?
4. Do exists several and significance differences between countries in energy sector and how the enterprises play their roles?
5. Trying to focus in the Spanish sector, can you please check my purpose of the key factors that can provide a full understanding of what a product is in the energy sector, and of course, improve that?
6. Can you please provide a summary about the product generation process in a marketer energy enterprise?
7. What are in your experience what do evaluate more the customers in these products?
8. Currently, do you evaluate customer feedback to customize the products? In this case how do you do it? (Automatic, business managers, etc...)
9. Do exists products restrictions applied for the government? Exist several significances in product factors between marketers companies or do all play with the same cards?(this allows to know if this algorithm is interesting to customize or is probably reliable for any enterprise)

Interview to Maria Macià – SAP Functional Consultant at ViewNext in the energy sector

1. First of all, can you introduce yourself? (A brief explanation of your studies and your professional career).

I studied Energy Engineering in UPC. It was the first year for this degree because it appeared with the new Bologna process. The two first years of the career was common with all the other engineers (Mechanical, Electrical, Electronic...) and in the third course I took Energy specific subjects like Energy Sector Planning, Energy Transmission and Distribution, Control of Energy Systems, Energy Storage, Energy Integration, etc.

During my career I did a practicum in Arbora&Ausonia (from P&G). It was not related with energy sector, was about stock management.

On my Bachelor's Thesis, I did an Erasmus program in Munich in Universität der Bundeswehr München designing and building a demo model of a Dish Stirling Solar Power Plant (with sun tracking).

After that, I did a Master's degree in Energy Engineering specialized in Management Energy. *It's an international master's degree in Energy Engineering that deals with current energy problems from different perspectives: resources, production technologies, transport and distribution of energy, environmental impact, efficiency, saving and rational use, etc. The Master provides the knowledge and skills necessary to analyse case studies and manage projects on the generation, transformation, distribution and consumption of different types of energy.*

During the master I was working in HEWLETT-PACKARD as a Procurement Engineer Intern and then change to another department to work as a Business Analyst.

I want to change and start my professional career on energy sector, so I got a job in FACTOR ENERGIA assigned to the Technical Team *as a Key Account management, Energy Regulator Advisor: working and cooperating with private customer and public administrations, analysing the evolution of Electric Spanish Market, developing and implementation of new sales tools, products and strategies and cooperating with the business plan of the expansion to South America: study and analysis of Mexico Electric Market and regulation and legislation.*

Recently I got a new job in VIEWNEXT (AN IBM SUBSIDIARY) as a SAP Functional Consultant assigned to Iberdrola Scottish Power project.

2. Do you know about machine learning and artificial intelligence? (what is it, its applications, etc...)

I don’t know very much, the information I have is from some articles or books. It is a field of the artificial intelligence and it is related to develop algorithms where the machine can learn and develop some “knowledge”.

It can be applied in many fields and sectors like in medicine, learning, prediction, energy efficiency…

3. Do you know about automatic product generation algorithms? In case of a positive answer, do you know if any energy sector enterprise is trying to develop an algorithm with this purpose?

Have you been involved in?

I did not work with algorithms, but I am currently working in the creation of a tool that will help in the product generation process. The goal is to achieve a more automatically process for the worker and a more custom product for the customer.

4. Do exists several and significance differences between countries in energy sector and how the enterprises play their roles?

Yes, there are some differences.

Each country has its own Energy Market where they sell and buy the electricity. For example, in Spain it is OMIE, the Nord Pool for Nordic, Baltic, UK and German markets, Epex Spot for France, GME market in Italy… The prices can vary a lot between the countries and there are different factors that can influence on this like the demand, the resources, technical infrastructures, the international connections and also a political influence.

There are some projects related to design a new European electricity market, following new trends with the goal to align and integrate all the markets.

5. Trying to focus in the Spanish sector, can you please check my purpose of the key factors that can provide a full understanding of what a product is in the energy sector, and of course, improve that?

In the energy sector a product is basically a description about the customer. This description allows to know what the profile and trends of the customer is, and then the company can buy the expected energy that this customer will use. With this prediction the company add the profits and the cost for other associated services, and finally get an associated product.

Then, to get a correct definition of the customer we should know:

1. Customer consumption
2. Hired power
3. Product type (Fixed or indexed)
4. Postal code
5. Payment type: Direct debit or bank
6. Mailing bill
7. Electronic billing: Only applied to public administration. Use XML to help with governmental accountability.
8. Other electrical services: Electrical car, Smart home (not already in Spain)
9. Permanence commitment.
10. Dual contract: Gas and electricity in some contract
11. Maintenance services: Companies provide maintenance services for boilers, and others.
12. Green power: Certificate that shows the percentage of the energy that provides from renewable energies.
13. Energetical advisor: The Company provides an advisor to the customer. This will check the bill every X month, and check if is possible to save money, etc…

As I can see on your proposal, there are other factors that also are important. For example, number of members in the family, and another social variable that helps in the product definition, but all these factors are included in the above variables. For example, and increased consumption shows that there are or more members in the family or it is a big house.

Once you know all these characteristics, you check the fix costs and the variables costs for the tariff and then add the expected profits (€/KwH), and a discount applied over the final price.

Do exists business restrictions for product generation?

There is some restriction in the product generation process, but most of the time there are defined for each company, so all of them are different as different are the company. Only two rules are defined by the law. The customer typology is divided in tariffs.

These tariffs are related to consumption and customer power. Despite that, the product itself is not defined for any other stakeholder than the company.

The other political rule is that exists some special tariffs. They are known as TUR (Last resource tariffs). Only Endesa, Iberdrola, Gas Natural, Naturgas, EON, Madrileña de gas (COR) are obligated to provide for these tariffs. These tariffs are those which customers with a very low consumption. The government defined a volunteer price for the small consumer (PVPC) which is the price that COR can charge to the customers under these tariffs. This business model will only apply to this project if you expect to introduce the software to COR companies.

An example of business rule is to apply only the energetical advisor to huge customers, this means customers with tariffs 3.1 or 6.1.

Another one, is to sell indexed product to customers with tariffs over 3.0.

7. What are in your experience what do evaluate more the customers in these products?

The most important factor for the customers is the price. They are looking for the cheapest product and a product that guarantees no changes during a period of time. They want to know how much they are going to pay in the next “X” months. Even they know that this depends on their consumption.

However, there is a new trend where the customers are more focus on the personal attention of the retailer. They want to be informed and want to know what they are exactly paying for. Since now, lot of people didn’t know how to read an electric or gas bill, so, nowadays, people are more interested on knowing all the items there are in the bill and have full attention if they required or have any questions.

8. Currently, do you evaluate customer feedback to customize the products? In this case how do you do it? (Automatic, business managers, etc...)

I never worked with customer feedback directly. Despite that, I have seen some methodologies applied to gather customer feedback.

For example, some feasible options are:

1. The energetical advisor: This advisor stablishes direct communications with the customer, so it can get a very good feedback.
2. Call center: Call center receives and issue phone calls that also take profit of direct communications.
3. Customer surveys: Through website or mail marketing the company can get feedback asking some specific questions about their products and services.