Universitat Oberta de Catalunya

Predicción de errores en producción industrial de piezas, mediante clasificación supervisada con desbalanceo de clases.

José Ahias Lopez Portillo

Trabajo Final de Máster

Máster Universitario en Ciencia de Datos

Índice

- Introducción
- Estado del arte
- Presentación del problema y análisis descriptivo
- Análisis predictivo
- Conclusiones y trabajo futuro

Introducción

Motivación

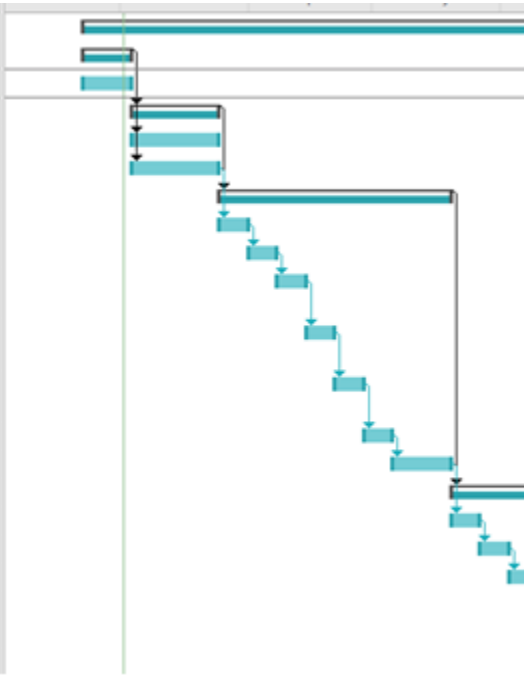
- Resolver un problema de clasificación de clases desbalanceadas.
- Usar datos de un entorno real de producción.
- Utilizar datos generados por sensores.
- Generar un modelo de aprendizaje automático de clasificación con clases desbalanceadas.

Objetivo

Crear un modelo de clasificación para predecir cuándo una pieza producida en línea no superara el control de calidad. Utilizando algoritmos de aprendizaje automático aplicados sobre el conjunto de datos de Kaggle y la fabrica de Bosch.

Planificación

Predicción errores en producción industrial	**84 days**	**Wed 2/20/19**	**Sun 6/16/19**	
Definición y planificación	**9 days**	**Wed 2/20/19**	**Sun 3/3/19**	
Avance de memoria de TFM	9 days	Wed 2/20/19	Sun 3/3/19	
Análisis de mercado	**16 days**	**Mon 3/4/19**	**Sun 3/24/19**	**2**
Recopilación de información y bibliografía	16 days	Mon 3/4/19	Sun 3/24/19	2
Avance de memoria de TFM	16 days	Mon 3/4/19	Sun 3/24/19	2
Diseño e implementación	**41 days**	**Mon 3/25/19**	**Sun 5/19/19**	**4**
Instalación y configuración de ambiente de desarrollo	6 days	Mon 3/25/19	Sun 3/31/19	6
Análisis exploratorio y preprocesamiento	6 days	Mon 4/1/19	Sun 4/7/19	8
Iteración 1 -> Implementación de algoritmos de Machine Learning	6 days	Mon 4/8/19	Sun 4/14/19	9
Iteración 2 -> Implementación de algoritmos de Machine Learning	6 days	Mon 4/15/19	Sun 4/21/19	10
Iteración 3 -> Implementación de algoritmos de Machine Learning	6 days	Mon 4/22/19	Sun 4/28/19	11
Análisis de resultados	6 days	Mon 4/29/19	Sun 5/5/19	12
Avance de memoria de TFM	11 days	Mon 5/6/19	Sun 5/19/19	13
Redacción de la memoria	**16 days**	**Mon 5/20/19**	**Sun 6/9/19**	**7**
Iteración 1 – Documento Final	6 days	Mon 5/20/19	Sun 5/26/19	14
Iteración 2 – Documento Final	6 days	Mon 5/27/19	Sun 6/2/19	16
Documento Final	6 days	Mon 6/3/19	Sun 6/9/19	17
Presentación y defensa	6 days	Mon 6/10/19	Sun 6/16/19	15



Estado del arte

Estado del arte

El desbalanceo de clases es una característica o error que se genera cuando al menos una o más clases (clases minoritarias) se encuentra representada significativamente en una menor cantidad con respecto a las otras (clases mayoritarias).

Estado del arte

- Detección de fraude en llamadas telefónicas. (Fawcett, Tom. y Provost, Foster.: Adaptive Fraud Detection. Data Minining and Knowledge Discovery, 1997. pp 1-28)

- Diagnósticos de fallos en equipos de telecomunicaciones (Weiss, Gary M. Hrish, Haym. Learning to predict rare events in event sequences. Appears in Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. 1998. pp 359-363).

Estado del arte

- En finanzas (Korn Sue.The Opportunity for Predictive Analytics in Finance. 2011)

- Prevención de intrusiones (De la Hoz Emiro. De la Hoz Eduardo., Ortiz Andrés. Ortega Julio. Modelo de detección de intrusiones en sistemas de red, realizando selección de características con FDR y entrenamiento y clasificación con SOM. 2012.)

Presentación del problemas y análisis descriptivo

Presentación del problemas

- Conjunto de datos con los datos de producción de Bosch.
- Problemas de competición de Kaggle.
- Más de un millón de observaciones.
- 970 dimensiones.
- Clase positiva (0 = piezas sin defectos).
- Clase negativa (1 = piezas con defectos).

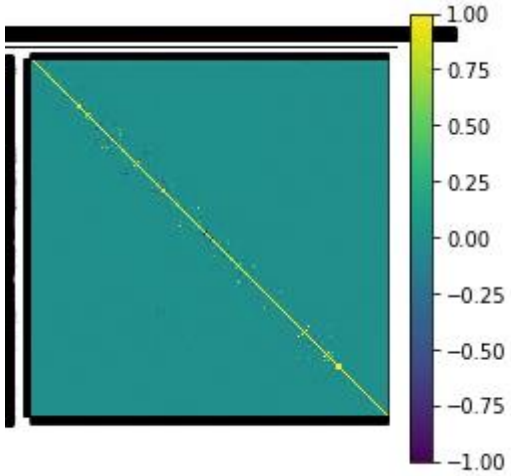
Análisis descriptivo

Análisis descriptivo

	per_nan
count	968.000000
mean	81.084969
std	30.681883
min	5.350000
25%	80.940000
50%	95.330000
75%	98.990000
max	99.890000

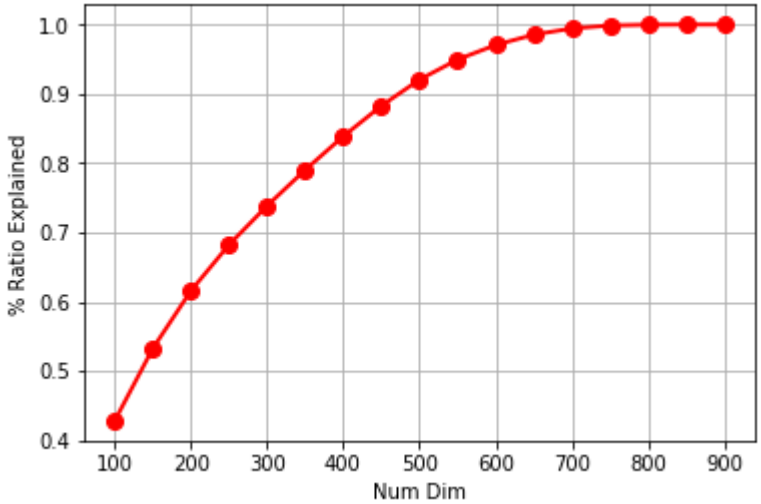
Un alto número de valores no definidos.
La media como imputación de valores.

Análisis descriptivo



Baja correlación entre dimensiones.

Análisis descriptivo



PCA (650 dimensiones)

Análisis predictivo

Experimentos

Experimento	Entrenamiento (Train)	Pruebas (Test)
1	169576	72677
2	251957	107982

Experimentos

Técnicas de muestreo:

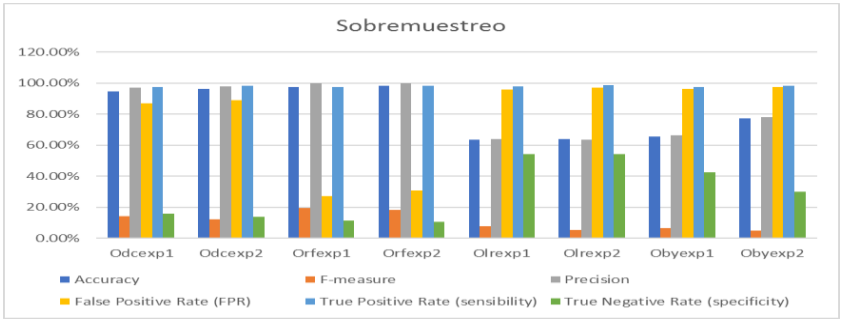
- Sobremuestreo.
- Submuestreo.
- Sobremuestreo sintético (SMOTE).

Aprendizaje automático:

- árboles de decisiones.
- Bosques aleatorios.
- Regresión logística.
- Clasificadores de bayes
- Máquina de soporte vectorial de una clase.

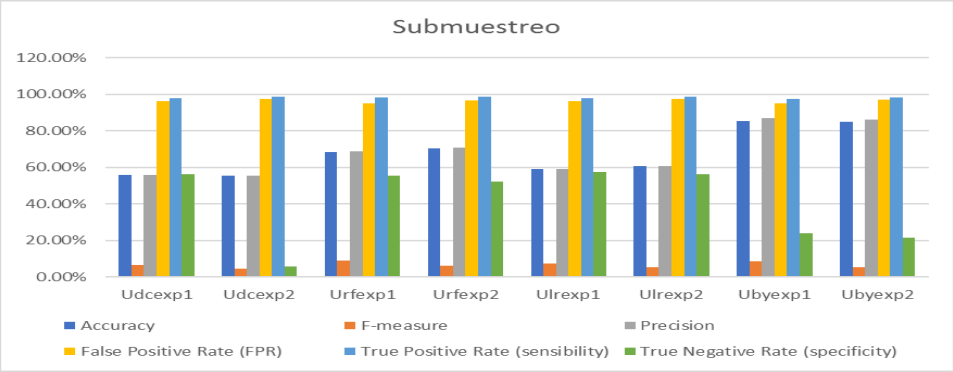
Resultados - Sobremuestreo

Métricas	Odcexp1	Odcexp2	Orfexp1	Orfexp2	Olrexp1	Olrexp2	Obyexp1	Obyexp2
Accuracy	94.60%	96.21%	97.36%	**98.20%**	63.56%	63.94%	65.65%	77.25%
F-measure	14.30%	12.18%	19.69%	**18.26%**	7.80%	5.45%	6.56%	4.83%
Precision	96.91%	97.81%	99.88%	**99.91%**	63.83%	63.73%	66.32%	78.17%
False Positive Rate (FPR)	86.98%	89.08%	27.24%	**30.67%**	95.80%	97.13%	96.45%	97.38%
True Positive Rate (sensibility)	97.52%	98.31%	97.47%	**98.28%**	97.95%	98.63%	97.53%	98.29%
True Negative Rate (specificity)	15.84%	13.76%	11.39%	**10.51%**	54.31%	54.36%	42.44%	30.18%



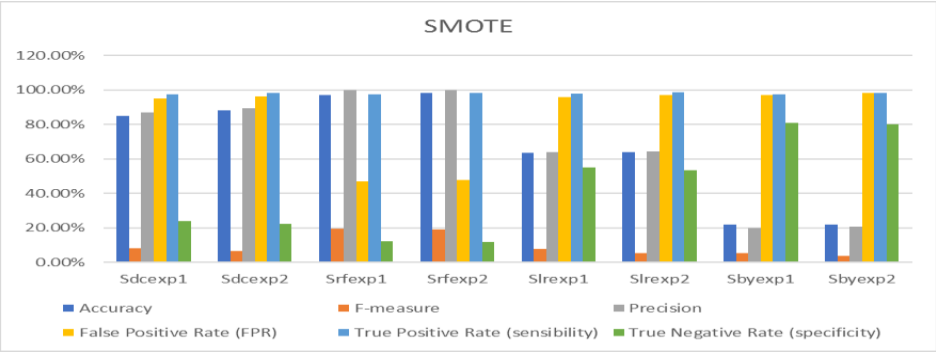
Resultados - Submuestreo

Métricas	Udcexp1	Udcexp2	Urfexp1	Urfexp2	Ulrexp1	Ulrexp2	Ubyexp1	Ubyexp2
Accuracy	55.70%	55.32%	68.38%	70.45%	59.00%	60.78%	**85.30%**	84.99%
F-measure	6.74%	4.70%	9.08%	6.32%	7.35%	5.20%	**8.46%**	5.19%
Precision	55.68%	55.27%	68.76%	70.81%	59.05%	60.87%	**87.09%**	86.22%
False Positive Rate (FPR)	96.41%	97.55%	95.05%	96.64%	96.07%	97.28%	**94.86%**	97.05%
True Positive Rate (sensibility)	97.76%	98.53%	98.15%	98.70%	97.93%	98.62%	**97.51%**	98.26%
True Negative Rate (specificity)	56.40%	5.77%	55.62%	52.13%	57.27%	56.25%	**23.93%**	21.51%



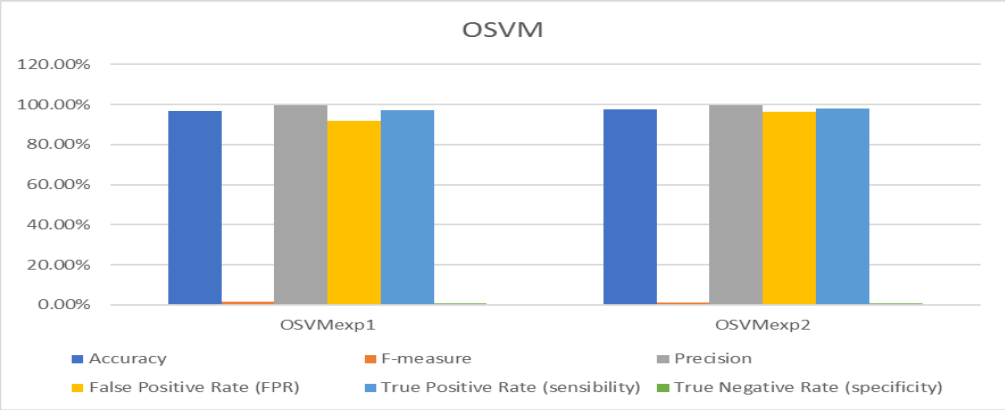
Resultados - SMOTE

Métricas	Sdcexp1	Sdcexp2	Srfexp1	Srfexp2	Slrexp1	Slrexp2	Sbyexp1	Sbyexp2
Accuracy	85.04%	88.20%	97.20%	**98.11%**	63.51%	63.95%	21.78%	21.95%
F-measure	8.36%	6.69%	19.58%	**19.22%**	7.87%	5.36%	5.54%	3.78%
Precision	86.82%	89.48%	99.69%	**99.79%**	63.77%	64.16%	20.06%	20.82%
False Positive Rate (FPR)	94.94%	96.06%	47.12%	**47.74%**	95.76%	97.18%	97.13%	98.07%
True Positive Rate (sensibility)	97.51%	98.33%	97.49%	**98.31%**	97.97%	98.60%	97.32%	98.17%
True Negative Rate (specificity)	24.03%	22.14%	12.02%	**11.77%**	54.89%	53.39%	80.72%	80.14%



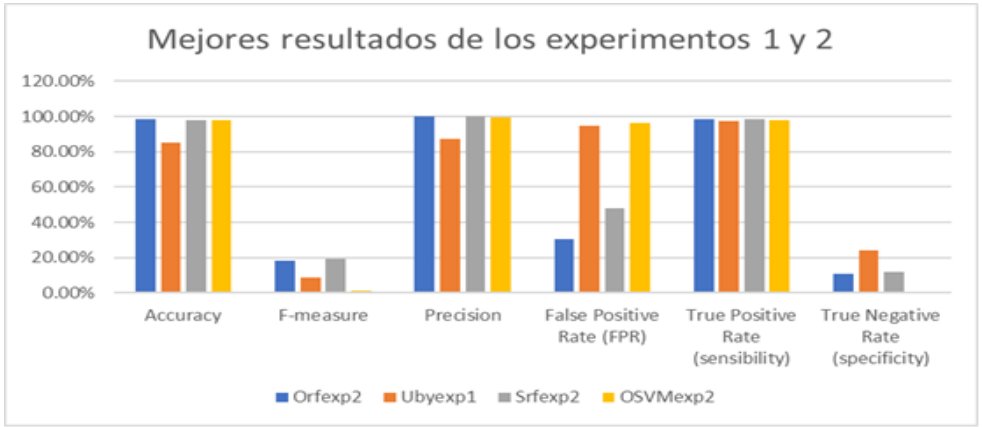
Resultados - OSVM

Métricas	OSVMexp1	OSVMexp2
Accuracy	96.90%	**97.70%**
F-measure	1.66%	**1.27%**
Precision	99.70%	**99.59%**
False Positive Rate (FPR)	91.67%	**96.42%**
True Positive Rate (sensibility)	97.18%	**98.10%**
True Negative Rate (specificity)	0.92%	**0.78%**



Resultados – Mejores modelos

Métricas	Orfexp2	Ubyexp1	Srfexp2	OSVMexp2
Accuracy	**98.20%**	85.30%	**98.11%**	97.70%
F-measure	**18.26%**	8.46%	**19.22%**	1.27%
Precision	**99.91%**	87.09%	**99.79%**	99.59%
False Positive Rate (FPR)	**30.67%**	94.86%	**47.74%**	96.42%
True Positive Rate (sensibility)	**98.28%**	97.51%	**98.31%**	98.10%
True Negative Rate (specificity)	**10.51%**	23.93%	**11.77%**	0.78%



Conclusiones y trabajo futuro

Conclusiones

- Existe una diferencia marcada entre el mundo académico y el mundo real empresarial.
- Las infraestructuras en la Nube pueden solucionar problemas de recursos, pero es costoso económicamente.
- De los 26 experimentos sobre el conjunto de datos, los mejores resultados se obtienen con: sobremuestreo, SMOTE y bosques aleatorios.

Trabajo futuro

- Nuevos experimentos con más observaciones enfocados a utilizar técnicas de sobremuestreo, SMOTE y bosques aleatorios.

Trabajo futuro

- Utilizar parámetros aleatorios para mejorar la configuración y el rendimiento del modelo de bosque aleatorios.
- Probar nuevos algoritmos como **“Isolation Forest**” o **“redes neuronales”**