upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Trabajo Final
Online Shopper's Intention

Hanwei, ke: https://github.com/Sendaia

Jhonny Fabricio, Chicaiza Palomo: https://github.com/jhonnyfc

https://github.com/jhonnyfc/Online_Shoppers_Intention

13/05/20120

Índice

Problema seleccionado	3
Preparación de datos	4
Selección de parámetros	5
Elección del Modelo	6
Conclusiones	6

Presentación: https://prezi.com/mdqeg7aho4bm/online_shoppers_intention/

Problema seleccionado

Después de seleccionar varios problemas hemos elegido finalmente el data set de Online Shopper's Intention (https://www.kaggle.com/roshansharma/online-shoppers-intention). Se elegido este debido a que es un tema actual y es de gran interés para las empresas online emergentes o para las ya establecidas. El problema a resolver será averiguar si un usuario va a realizar o no una transacción online.

El data set consta de 12316 ejemplos y 17 variables, de las que 14 son numéricas y 3 categóricas:

Numéricas:

- 'Administrative': Número de páginas de tipo administrativo
- 'Administrative_Duration': mins, duración en este tipo de paginas
- 'Informational': Número de páginas de tipo información
- 'Informational_Duration': mis, duración en este tipo de paginas
- 'ProductRelated': Número de páginas de productos relacionados vistos
- 'ProductRelated_Duration': mins, tiempo que ha estado en las páginas de productos relacionados
- 'BounceRates': (0,1) Porcentaje de visitante que entran y salen sin interactuar
- 'ExitRates': (0,1) Porcentaje de persona que salieron después de interactuar
- 'PageValues': Media de la valoración de la página antes de hacer una transacción comercial
- 'SpecialDay': (0,1) Si el día está cerca de un día especial
- 'OperatingSystems': Este valor nos dice que sistema operativo usa el usuario, no nos aporta mucha información debido a que los valores que tomaban eran enteros.
- 'Browser': Este valor nos da el tipo del buscador que utiliza el usuario, pero al ser un número no podemos saber que buscador es.
- 'Region': Este valor nos da la región que pertenece, pero al ser un número no podemos sacar conclusiones,
- 'TrafficType': Este valor nos da el tipo de tráfico de la web por que se ha llegado, pero debido a que es numérica no podemos sacar conclusiones.

Categóricas:

- 'Month': 'Feb': 2,'Mar': 3,'May': 5,'June': 6,'Jul': 7,'Aug': 8,'Sep': 9,'Oct': 10,'Nov': 11,'Dec': 12
- 'Weekend': 1(False), 2(True)
- 'VisitorType': 1(Ha vuelto), 2(Nuevo), 3(otro)

Tras analizar el significado de cada uno de los datos hemos obtenido que la variable 'PageValues' es de gran importancia para la clasificación del data set.

Preparación de Datos

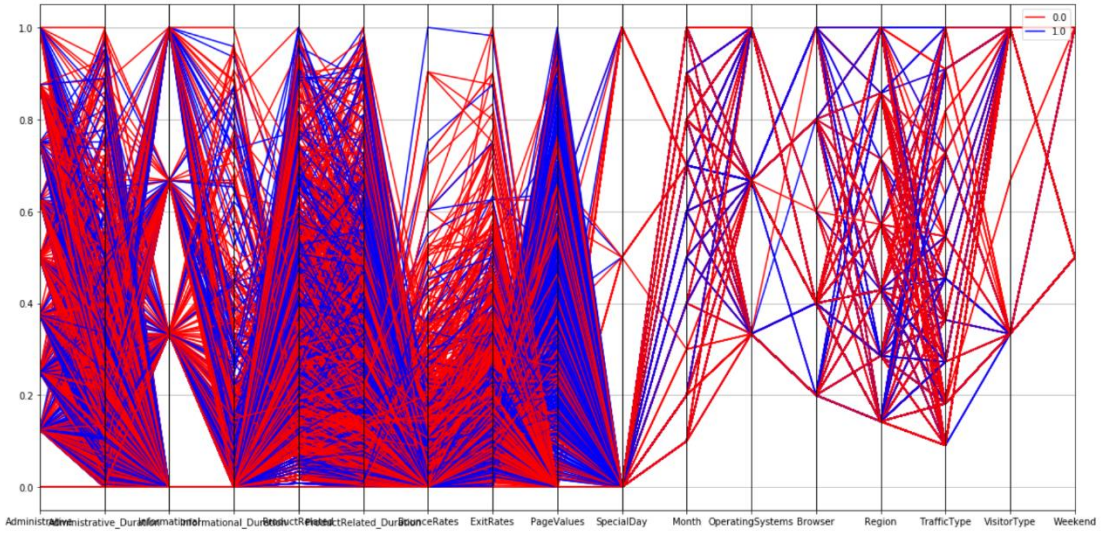
En el data set nos encontramos con 1908 ejemplos de clase positiva y 10408 de clase negativa.

Este data se consta además con 112 **valores perdidos** que pertenecen a 14 ejemplos, hemos decidido quitar estos ejemplos debido a la gran cantidad de ejemplos.

El siguiente factor para analizar es la **detección de Outliers** en la variable numéricas. Con el valor k = 2 se han detectado 5715 filas con Outliers por lo que hemos decidido eliminarlos, quedándonos con 6601 ejemplos, de los cuales 820 son de la clase positiva y 5781 son de la clase negativa.

Se ha decidido hacer **selección de instancias**, para ello hemos utilizado CNN y RNN para analizar las técnicas hemos dividido el conjunto anterior en test y train. Con CNN hemos obtenido un subconjunto de 1472 con un accuracy en train del 84% y en test del 87% frente al 82% en train y el 87% en test de RNN con un subconjunto de menor. Por lo que se ha decidido utilizar el RNN sobre el conjunto del paso anterior obteniendo así un conjunto de 1620 ejemplos, donde 595 son de la calase positiva y 1085 son de la clase negativa.

Tenemos 17 variables en el data set. En la investigación previa de cada una de esta variable pensamos que hay algunas que se podían quitar por lo que hemos decido realizar **selección de variables**. Mediante la visualización de la matriz de coordenadas paralelas podemos observar que donde menos ruido hay es en la variable 'PageValues'.



Por otra parte, hemos seleccionado 4 calificadores RandomForest, Decision Tree, SVM y KNN, para los tres primeros hemos selecciona las variables más importantes de cada modelo. Donde vemos que en todas sus listas sale la variable 'PageValues'.

Selección de Parametros

Hemos seleccionado 4 calificadores RandomForest, Decision Tree, SVM y KNN, como para los tres primero hemos obtenido sus mejores variables. Vamos a realizar el GridSearchCV con las variables seleccionadas para cada clasificador, para el KNN le pasamos el conjunto previo a la selección de variables.

Como estimador de calidad el accuracy, aunque hemos probado fscore y otros.

Para Random Forest Hemos obtenido:

> criterion= 'gini', max_depth= 5, max_features= 'log2', min_samples_split= 20, n_estimators= 10

Para Decision Tree hemo obtenido:

> criterion= 'gini', min_samples_leaf= 4, min_samples_split= 15

Para SVM hemos obtenido:

> C= 1

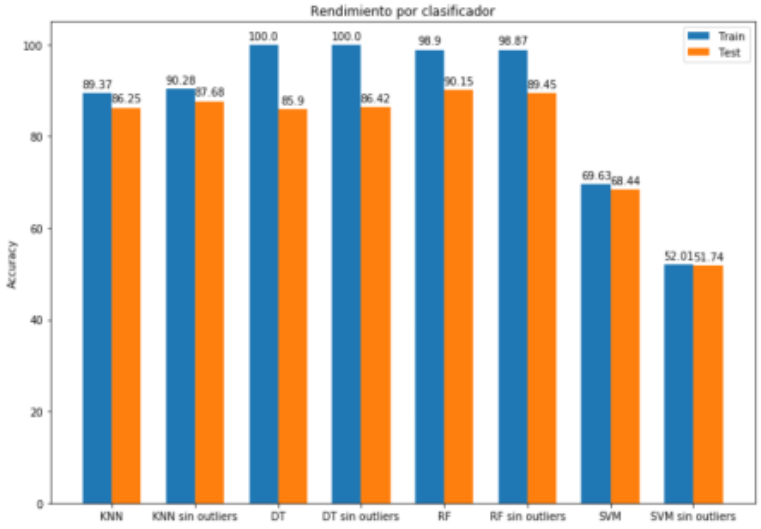
Para KNN hemos obtenido:

> n_neighbors= 9, p= 1, weights= 'uniform'

El proceso más costos para obtener estos valores ha sido el de del random Forest.

Elección del Modelo

Primero hemos entrenado los 4 clasificadores nombrados anteriormente con sus valores por defecto y con los datos sin procesar lo hemos comparado con los resultados de los clasificadores entrenados con los datos sin Outliers. Se han obtenido unos valores muy parecidos de accuracy.



Por otra parte, con los valores obtenidos en selección de parámetros hemos entrenado los 4 clasificadores con los datos procesados que tiene una dimensión de 1620 y para probar su calidad hemos utilizado el conjunto original.

KNN: El rendimiento con todas las variables es el 85.66%

	precision	recall	f1-score
0	0.87	0.98	0.92
1	0.62	0.20	0.30

Decision Tree: El rendimiento con todas las variables es el 83.08%

	precision	recall	f1-score
0	0.94	0.86	0.90
1	0.47	0.68	0.56

Random Forest: El rendimiento con todas las variables es el 83.92%

	precision	recall	f1-score
0	0.94	0.86	0.90
1	0.49	0.72	0.58

SVM Lineal: El rendimiento con todas las variables es el 84.37%

	precision	recall	f1-score
0	0.85	1.00	0.92
1	0.19	0.00	0.01

Como podemos ver el que peor se comporta con respecto a las medidas de calidad es el SVM, aunque este no tenga un mal valor en accuracy con respecto el entrenamiento inicial.

Los demás clasificadores tienen unos valores de calidad más equilibrados. Pero el que mejor media tiene es el Random Forest, aunque el que ha obtenido el mejor accuracy sobre el conjunto origina es el clasificador KNN.

Por lo que en la práctica se podría utilizar cualquiera de los dos modelos, aunque se recomendaría utilizar ReandomForest

Queda destacar que también probamos a entrenar un clasificador solo con la variable 'PageValues' obteniendo un 89% de accuracy en test.

Conclusiones

Se puede ver que reflejado en las variables seleccionadas que una persona que busca productos relacionados, que pase un tiempo considerable para la variable ProductRelated, que tenga un BouneRate bajo y un PageValues alto va a comprar seguro un producto en la tienda visitada.