Inteligencia de Negocio (2017-2018)
Grado en Ingeniería Informática
Universidad de Granada

Práctica 3

Juan José Sierra González
jjsierra103@gmail.com

7 de enero de 2018

Índice

Índice de figuras

Índice de tablas

1. Introducción

La última práctica de la asignatura de Inteligencia de Negocio consiste en participar en una competición dentro de la conocida plataforma de ciencia de datos Kaggle. Esta competición se realiza a nivel mundial pero los alumnos de la UGR competimos entre nosotros para ver quién consigue la mejor solución, es decir, obtener un error cuadrático medio menor sobre las predicciones de la variable respuesta. Como la competición es de la categoría "Getting started" dentro de la plataforma está indicado que es una competición orientada al aprendizaje, y se pueden encontrar muchos kernels y manuales de ayuda para facilitar la comprensión y el estudio del problema.

El problema al que nos enfrentamos es tratar de predecir el precio de aproximadamente 1500 viviendas de Ames, Iowa, a partir de otras 1500 viviendas de las que conocemos 79 variables descriptivas y una variable respuesta (el precio de venta de la casa) que será la que tendremos que predecir.

2. Tabla de resultados

En esta competición he realizado un total de **18 subidas a Kaggle**. Mi posición final fue la 456, con un error cuadrático medio de 0.11692. A continuación se mostrará una tabla con los resultados obtenidos en cada una de las subidas y una pequeña descripción del modelo empleado. Así se podrá observar de forma clara la evolución que se ha ido produciendo en los modelos y cómo esto ha influido en el resultado obtenido.

Subida	Posición	Score	Fecha	Hora	Train RMSLE	Preprocesado	Algoritmos utilizados y parámetros
1	1049	0.12611	28/12/2017	17:30	-	Eliminación características >50 % NA y no correladas, llenado de NA en train, logaritmo etiquetas, dummies	ElasticNet(alpha=[0.0001..10], l1ratio=[0.01..0.99]) y GradientBoosting (estimadores=3000, learning_rate=0.05, max_depth=3)
2	1031	0.12559	31/12/2017	13:50	-	Eliminación características >50 % NA y no correladas, llenado de NA en train, logaritmo etiquetas, dummies	ElasticNet(alpha=[0.0001..10], l1ratio=[0.01..0.99]), GradientBoosting (estimadores=3000, learning_rate=0.05, max_depth=3) y XGBoost (estimadores=3000, learning_rate=0.05, max_depth=3)
3	1031	0.13545	31/12/2017	13:59	-	Eliminación características >50 % NA y no correladas, llenado de NA en train, logaritmo etiquetas, dummies	XGBoost (estimadores=3000, learning_rate=0.05, max_depth=3)
4	1016	0.13044	02/01/2018	12:00	-	Filtrado características con muchos NA y con información duplicada y no correladas, llenado de NA en train y test logaritmo etiquetas, dummies	ElasticNet(alpha=[0.0001..10], l1ratio=[0.01..0.99]), GradientBoosting (estimadores=3000, learning_rate=0.05, max_depth=3) y XGBoost (estimadores=3000, learning_rate=0.05, max_depth=3)

Tabla 2.1: Resultados descriptivos de cada una de las subidas a Kaggle.