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

Práctica 3

Juan José Sierra González
jjsierra103@gmail.com

8 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	ElasticNet: 0.11921 GBoosting: 0.07561	Eliminación características >50% NA y no correladas, imputación 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	ElasticNet: 0.11921 GBoosting: 0.07588 XGBoost: 0.02117	Eliminación características >50% NA y no correladas, imputación 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	XGBoost: 0.02117	Eliminación características >50% NA y no correladas, imputación 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	ElasticNet: 0.12619 GBoosting: 0.07604 XGBoost: 0.02307	Filtrado características con muchos NA y con información duplicada y no correladas, imputación 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)
5	1023	0.12832	02/01/2018	15:23	ElasticNet: 0.10232 GBoosting: 0.06869 XGBoost: 0.01971	Eliminación características >50% NA y no correladas, eliminación outliers, imputación de NA en train y test (utilizando mediana y moda), 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)
6	1024	0.12690	02/01/2018	16:39	ElasticNet: 0.09642 GBoosting: 0.06507 XGBoost: 0.01717	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), 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)
7	1039	0.12720	03/01/2018	15:05	ElasticNet: 0.09510 GBoosting: 0.10801 XGBoost: 0.06490	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), logaritmo etiquetas, dummies, transformación box-cox a todas las variables	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)
8	1041	0.13346	03/01/2018	16:14	ElasticNet: 0.09705	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), logaritmo etiquetas, dummies, transformación box-cox solo a variables muy sesgadas	ElasticNet(alpha=[0.0001..10], l1ratio=[0.01..0.99])
9	492	0.11789	03/01/2018	18:38	0.1088	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), logaritmo etiquetas, dummies, transformación box-cox solo a variables muy sesgadas, label encoding a algunas variables categóricas	Stacked Model con ElasticNet (alpha=0.0005, l1ratio=0.9), GradientBoosting (estimadores=3000, learning_rate=0.05, max_depth=4), Lasso (alpha=0.0005), y KernelRidge (alpha=0.6, grado=2, coef0=2.5)
10	502	0.11906	04/01/2018	9:33	0.1093	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), logaritmo etiquetas, dummies, transformación box-cox solo a variables muy sesgadas, label encoding a algunas variables categóricas	Stacked Model con ElasticNet (alpha=0.0005, l1ratio=0.9), XGBoost (estimadores=2200, learning_rate=0.05, max_depth=3), Lasso (alpha=0.0005), y KernelRidge (alpha=0.6, grado=2, coef0=2.5)
11	450	0.11712	04/01/2018	10:11	0.1085	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), logaritmo etiquetas, dummies, transformación box-cox solo a variables muy sesgadas, label encoding a algunas variables categóricas	Stacked Model con ElasticNet (alpha=0.0005, l1ratio=0.9), GradientBoosting (estimadores=3000, learning_rate=0.05, max_depth=4), XGBoost (estimadores=2200, learning_rate=0.05, max_depth=3), Lasso (alpha=0.0005), y KernelRidge (alpha=0.6, grado=2, coef0=2.5)

Tabla 2.1: Resultados descriptivos de cada una de las subidas a Kaggle (subidas 1-11).

Subida	Posición	Score	Fecha	Hora	Train RMSLE	Preprocesado	Algoritmos utilizados y parámetros
12	453	0.11722	04/01/2018	12:10	0.1078	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), logaritmo etiquetas, dummies, transformación box-cox solo a variables muy sesgadas, label encoding a algunas variables categóricas, ranking de valores en variables categóricas	Stacked Model con ElasticNet (alpha=0.0005, l1ratio=0.9), GradientBoosting (estimadores=3000, learning_rate=0.05, max_depth=4), XGBoost (estimadores=2200, learning_rate=0.05, max_depth=3), Lasso (alpha=0.0005), y KernelRidge (alpha=0.6, grado=2, coef0=2.5)
13	447	0.11692	04/01/2018	15:11	0.1075	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), logaritmo etiquetas, dummies, transformación box-cox solo a variables muy sesgadas, label encoding a algunas variables categóricas, ranking de valores en variables categóricas, simplificado del ranking y combinación de variables	Stacked Model con ElasticNet (alpha=0.0005, l1ratio=0.9), GradientBoosting (estimadores=3000, learning_rate=0.05, max_depth=4), XGBoost (estimadores=2200, learning_rate=0.05, max_depth=3), Lasso (alpha=0.0005), y KernelRidge (alpha=0.6, grado=2, coef0=2.5)
14	447	0.11793	04/01/2018	16:04	0.1073	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), logaritmo etiquetas, dummies, transformación box-cox solo a variables muy sesgadas, ranking de valores en variables categóricas, simplificado del ranking y combinación de variables, label encoding a nuevas variables categóricas, añadiendo las simplificadas	Stacked Model con ElasticNet (alpha=0.0005, l1ratio=0.9), GradientBoosting (estimadores=3000, learning_rate=0.05, max_depth=4), XGBoost (estimadores=2200, learning_rate=0.05, max_depth=3), Lasso (alpha=0.0005), y KernelRidge (alpha=0.6, grado=2, coef0=2.5)
15	451	0.11765	04/01/2018	17:25	0.1069	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), logaritmo etiquetas, dummies, transformación box-cox solo a variables muy sesgadas, ranking de valores en variables categóricas, simplificado del ranking y combinación de variables, label encoding a nuevas variables categóricas, añadiendo las simplificadas, transformaciones exponenciales de variables más correladas con la variable respuesta	Stacked Model con ElasticNet (alpha=0.0005, l1ratio=0.9), GradientBoosting (estimadores=3000, learning_rate=0.05, max_depth=4), XGBoost (estimadores=2200, learning_rate=0.05, max_depth=3), Lasso (alpha=0.0005), y KernelRidge (alpha=0.6, grado=2, coef0=2.5)
16	452	0.11774	04/01/2018	17:46	0.1069	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), logaritmo etiquetas, dummies, transformación box-cox solo a variables muy sesgadas, ranking de valores en variables categóricas, simplificado del ranking y combinación de variables, label encoding a anteriores variables categóricas, añadiendo las simplificadas, transformaciones exponenciales de variables más correladas con la variable respuesta	Stacked Model con ElasticNet (alpha=0.0005, l1ratio=0.9), GradientBoosting (estimadores=3000, learning_rate=0.05, max_depth=4), XGBoost (estimadores=2200, learning_rate=0.05, max_depth=3), Lasso (alpha=0.0005), y KernelRidge (alpha=0.6, grado=2, coef0=2.5)

Tabla 2.2: Resultados descriptivos de cada una de las subidas a Kaggle (subidas 12-16).

Subida	Posición	Score	Fecha	Hora	Train RMSLE	Preprocesado	Algoritmos utilizados y parámetros
16	452	0.11774	04/01/2018	17:46	0.1069	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), logaritmo etiquetas, dummies, transformación box-cox solo a variables muy sesgadas, ranking de valores en variables categóricas, simplificado del ranking y combinación de variables, label encoding a anteriores variables categóricas, añadiendo las simplificadas, transformaciones exponenciales de variables más correladas con la variable respuesta	Stacked Model con ElasticNet (alpha=0.0005, l1ratio=0.9), GradientBoosting (estimadores=3000, learning_rate=0.05, max_depth=4), XGBoost (estimadores=2200, learning_rate=0.05, max_depth=3), Lasso (alpha=0.0005), y KernelRidge (alpha=0.6, grado=2, coef0=2.5)
17	454	0.11830	04/01/2018	18:53	0.1070	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), logaritmo etiquetas, dummies, transformación box-cox solo a variables muy sesgadas, label encoding a anteriores variables categóricas, ranking de valores en variables categóricas, simplificado del ranking y combinación de variables	Stacked Model con ElasticNet (alpha=0.0002, l1ratio=0.9), GradientBoosting (estimadores=3000, learning_rate=0.02, max_depth=3), XGBoost (estimadores=2200, learning_rate=0.02, max_depth=3), Lasso (alpha=0.0005), y KernelRidge (alpha=0.6, grado=2, coef0=2.5)
18	456	0.11097	04/01/2018	19:25	0.1066	Eliminación outliers, imputación de NA en train y test para casi todas las variables (utilizando mediana y moda), logaritmo etiquetas, dummies, transformación box-cox solo a variables muy sesgadas, label encoding a anteriores variables categóricas, ranking de valores en variables categóricas, simplificado del ranking y combinación de variables	Stacked Model con ElasticNet (alpha=0.0002, l1ratio=0.9), GradientBoosting (estimadores=3000, learning_rate=0.02, max_depth=3) y Lasso (alpha=0.0005)

Tabla 2.3: Resultados descriptivos de cada una de las subidas a Kaggle (subidas 17-19).