Cat in the Dat

Informe Proyecto Final

Los datos que fueron escogidos por este grupo se recogen desde una competencia de Kaggle https://www.kaggle.com/c/cat-in-the-dat-ii/overview . El objetivo de esta competencia es predecir con probabilidad [0, 1] si un dato corresponde a un gato o no, vale decir, una variable de respuesta llamada "Target". La base de datos se encuentra dividida en "train" y "test" y cada una contiene 600,000 y 400,000 registros respectivamente y las variables que contiene son de tipo categóricas, nominales, ordinales y de tipo cíclicas (día y mes). Dentro de las descripciones de esta base no se encuentra la de sus variables, por lo que se desconoce a que corresponde cada variable.

Resultados del Análisis exploratorio.

Como resultado del análisis exploratorio pudimos determinar que, de las 26 variables que contiene la base, todas presentan datos faltantes exceptuando la variable target. Estos datos faltantes equivalían en el peor de los casos a un 3.048% del total. Además, pudimos determinar que las variables binarias bin_0, bin_1, bin_2 no se encuentran balanceadas. Para el caso de las nominales, pudimos notar que existen variables como nom_1, nom_2, nom_3 y nom_4 que contenían más de una categoría con frecuencia baja. Adicional a esto, notamos que la variable ord_5 parece contener un código alfabético, con 2218 categorías y una frecuencia máxima de 565. Por otro lado, el siguiente grupo de variables toma valores alfanuméricos con la cantidad de categorías y frecuencias señaladas:

Variable	categorías	frecuencia máxima
nom_5	1220	977
nom_6	1519	805
nom_7	222	5035
nom_8	222	5052
nom_9	2218	565

Como conclusión del análisis exploratorio, se decidió:

- imputar los valores faltantes por la moda de la variable. Esto ya que la base no contiene variables de tipo continua que nos permita pensar en otros tipos de imputación como el promedio y que la cantidad de datos faltantes es inferior a un 3.048%.
- Recodificar las variables nom_1, nom_2, nom_3 y nom_4 de la siguiente forma:
    - Nom_1 categoría "Square" y "Star" como "Others".
    - Nom_2 categoría "Cat" y "Sanke" como "Others".
    - Nom_3 categoría "Canada" y "China" como "Others".
    - Nom_4 categoría "Oboe" y "Piano" como "Others".
- eliminar las variables 'nom_5','nom_6', 'nom_7', 'nom_8', 'nom_9', 'ord_5' para no afectar la interpretabilidad de los modelos y los resultados.

Análisis exploratorio Univariado.

En este análisis consideramos la variable de respuesta 'target'. Pudimos observar que esta variable no se encuentra balanceada, presentando un 81.28% de valores 0 y un 18.72% de valores 1. Este problema fue

abordado con un submuestreo aleatorio. Esto dado que se cuenta con una gran cantidad de datos (600,000 registros) con lo cual minimizamos el efecto de perdida de información de esta técnica.

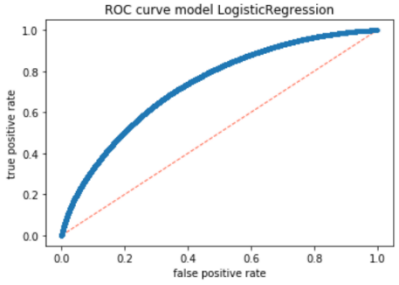
Recodificación de Variables

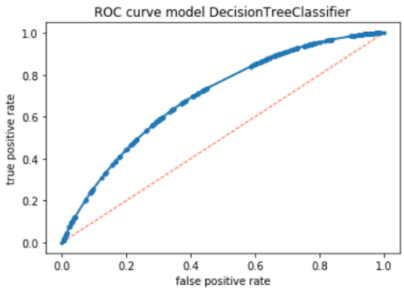
Se recodificaron las variables de acuerdo a lo siguiente:

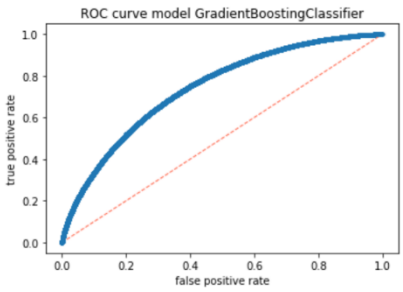
- ord_1 'Novice' = 1; 'Contributor'= 2; 'Expert'= 3; 'Master'= 4; 'Grandmaster'= 5,
- ord_2 'Freezing'= 1; 'Cold'= 2; 'Warm'= 3; 'Hot'= 4; 'Boiling Hot'= 5; 'Lava Hot'= 6,
- ord_3 toma valores desde la “a” a la “o” de forma consecutiva, por lo que se le asignó en el mismo orden, números del 1 al 15.
- ord_4 toma valores desde la “A” a la “Z” de forma consecutiva, por lo que se le asignó en el mismo orden, números del 1 al 26.
- Se transformó a *dummies* cada categoría de todas las variables que no eran binarias.

Modelamiento

Para este informe se abordaron predicciones con tres modelos, *Logit, Decision Tree* y *Gradient Boosting.* La metodología de trabajo consistió en entrenar los tres modelos con los datos, por medio de una grilla que nos permitió ajustar con los mejores hiperparámetros utilizando el 3 o 5 validaciones cruzadas. Finalmente, los resultados de los modelos fueron los siguientes:







	Modelo		
	LR	DTC	GBC
AUC	0.733	0.698	0.740
Accuracy	0.671	0.646	0.675
Presicion	0.671	0.645	0.675
Recall	0.671	0.646	0.675
F1	0.671	0.646	0.675

Podemos observar que en general, los 3 modelos tienen un ajuste bastante bueno (todos por sobre el 69% de AUC). La exactitud y precisión de los modelos también es bastante parecida, esto queda evidenciado en las curvas ROC anteriormente mostradas. Dado esto, preferiremos el modelo *Gradient Boostin Classifier* ya que en todos los indicadores presenta mejores resultados que los otros 2 modelos.