Minería de Datos

“

Nuestro objetivo es lograr
es predecir con probabilidad
[0, 1] si un dato corresponde a
un gato o no

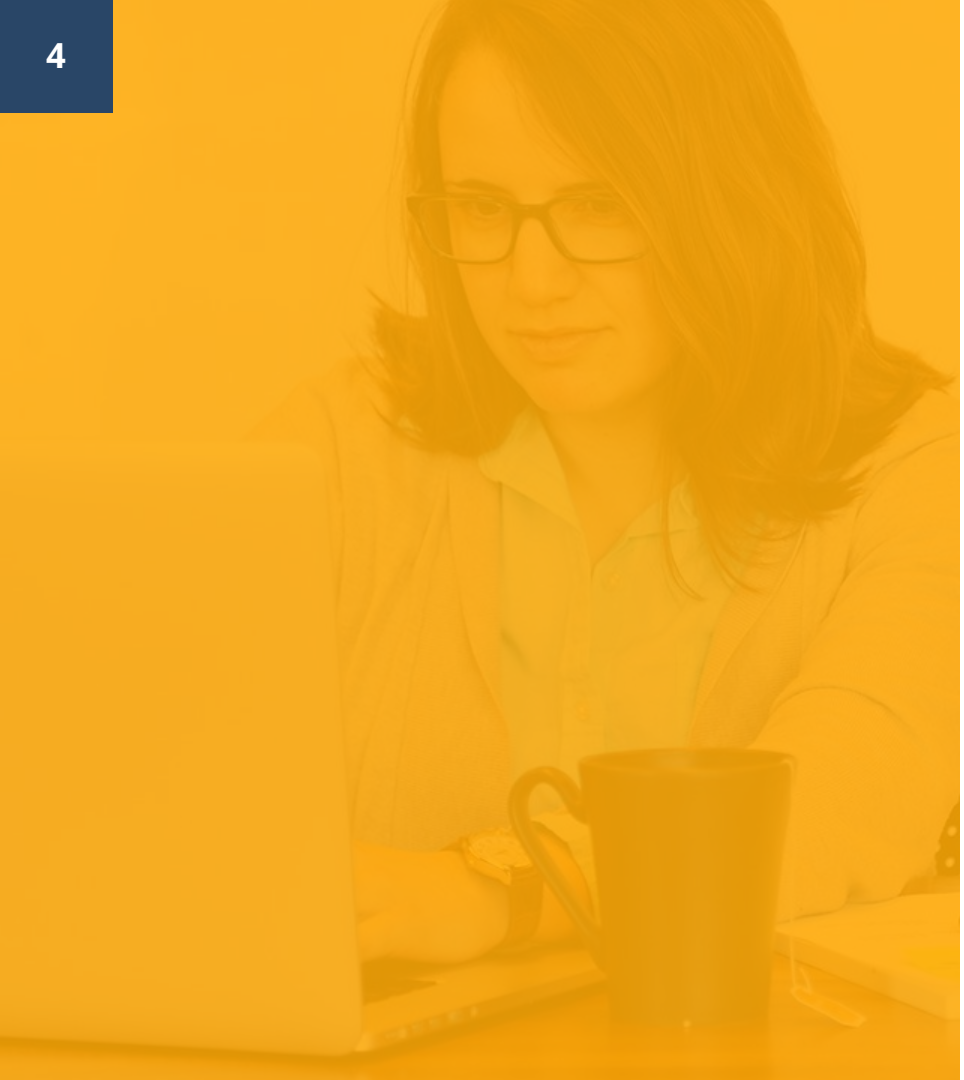
Descripción de la data

Esta data fue extraída de competencia de Kaggle, la cual se encuentra dividida en “train” y “test” , donde la variable predictora es llamada “target” y cada una contiene 600,000 y 400,000 registros respectivamente.

Las variables que contiene corresponden:

- 5 variables categoricas
- 10 variables nominales
- 6 variables ordinales
- 2 variables cíclicas (día y mes)

Cabes destacar que se desconoce la descripción de cada variable.



Análisis exploratorio

Análisis exploratorio

Exceptuando la variable target, todas las variables se encuentran con datos faltantes, donde el peor de los casos contiene 3.048%.

Estos datos faltantes se decidió imputarlos a la media de la variable.

También se tiene que las variables

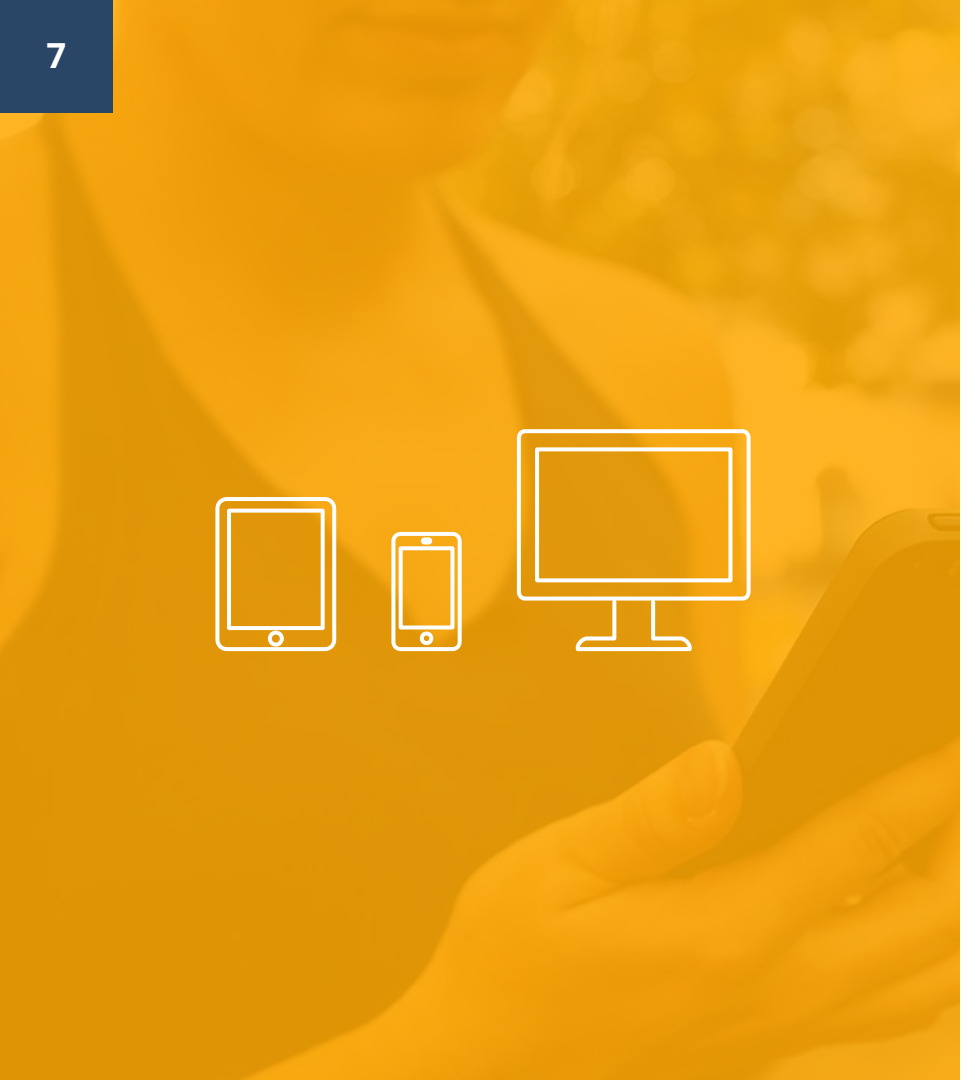
- bin_0, bin_1, bin_2 no se encuentran balanceadas.
- nom_1, nom_2, nom_3 y nom_4 que contenían más de una categoría con frecuencia baja
- la variable ord_5 parece contener un código alfabético, con 2218 categorías y una frecuencia máxima de 565, la cual se decidió eliminar, ya que no afecta interpretabilidad de los modelos y los resultados .

Análisis exploratorio

Las siguientes 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

Este conjunto de variables se decidió que debían ser eliminadas, ya que no afecta interpretabilidad de los modelos y los resultados



Análisis univariado

La variable target se encuentra desbalanceada, teniendo:

- 81.28% de valores 0
- 18.72% de valores 1

Como se tiene una gran cantidad de datos, este desbalance se trato con un submuestreo.

Recodificación de variables

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.

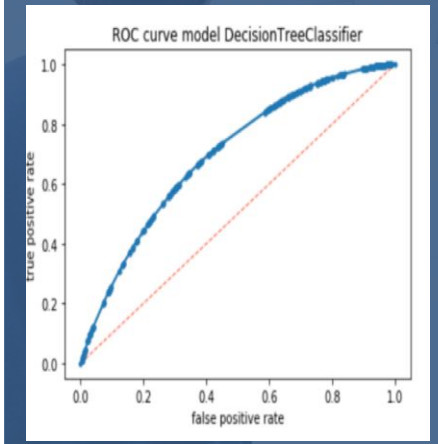
Modelamiento

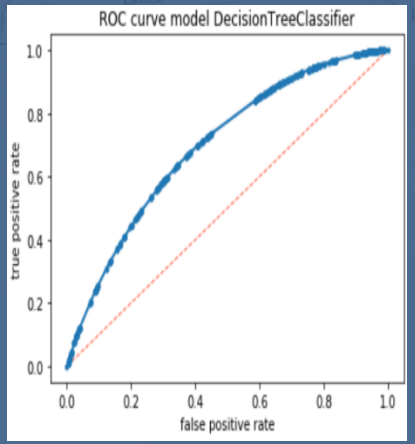
Modelos

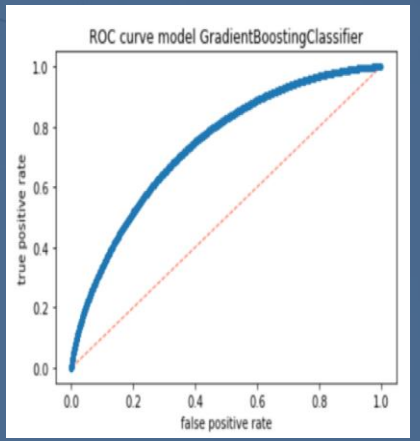
La metodología de trabajo consistió en entrenar tres modelos con los datos:

- *Logit*
- *Decision Tree*
- *Gradient Boosting*

Donde por medio de una grilla que nos permitió ajustar con los mejores hiperparámetros utilizando el 3 o 5 validaciones cruzadas.
A continuación les presentamos las curvas ROC de cada modelo:



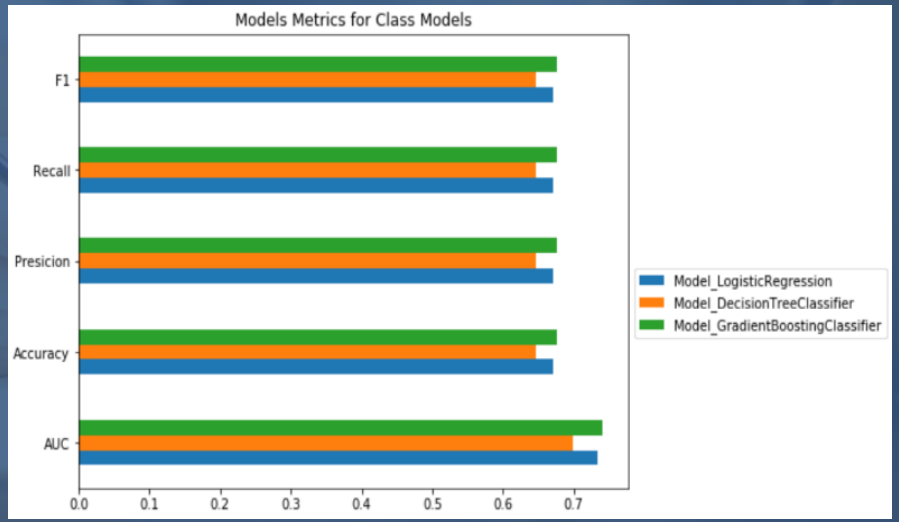




Resultados

	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

En base a lo ya mostrado preferiremos el modelo *Gradient Boostin Classifier* ya que en todos los indicadores presenta mejores resultados que los otros 2 modelos.



Gracias!

Preguntas?