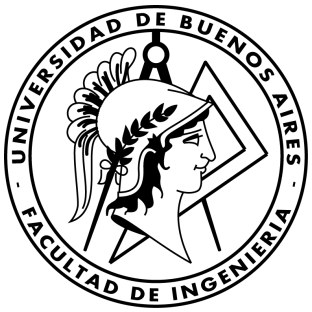

Trabajo Práctico 1
Reservas de hotel - Conclusiones generales

[75.06] Organización de datos
Primer cuatrimestre 2023
Grupo 19: Sudanalytics

Integrantes

Nombre	Padrón	Mail
Re, Adrián Leandro	105025	are@fi.uba.ar
Lorenzo, Luciano Andrés	108951	llorenzo@fi.uba.ar
Toulouse, Alan	105343	atoulouse@fi.uba.ar
Tonizzo, Nicolas	107820	ntonizzo@fi.uba.ar

Fecha de entrega: 25/05/2023

Índice

1. Introducción **2**

2. La importancia del feature engineering **2**

3. Modelos **3**

4. Conclusiones **3**

5. Referencias **3**

1. Introducción

El objetivo de este informe es dar una conclusión al trabajo practico en su totalidad describiendo sus aspectos mas relevantes como así también mencionar si quedaron opciones por explorar.

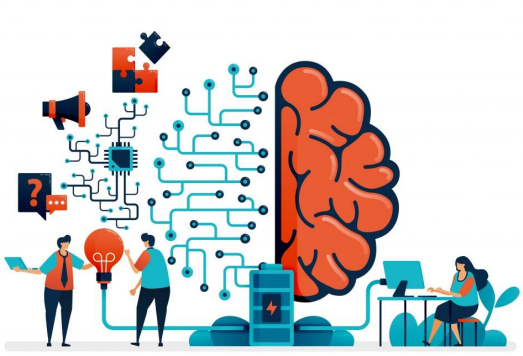

Figura 1: Ciencia de datos

2. La importancia del feature engineering

Creemos que la parte más importantes en nuestro trabajo fue la ingeniería de features durante el Checkpoint 1. Esta condicionaría los resultados de todos los Checkpoint siguientes. Principalmente la creación de nuevas variables ayudo en gran medida a nuestros modelos a encontrar patrones en los datos, que de otra forma no los hubiese encontrado.

Por ejemplo la feature 'es_extranjero', que describe si una reservación fue hecha por alguien portugués o no (los hoteles quedan en Portugal), es algo que los modelos nunca hubiesen podido tener en cuenta, porque no es algo explicito en el dataset que los hoteles sean portugueses.

Son este tipo de aspectos en los datos de los que los humanos con conocimientos sobre el dominio del problema podemos aprovecharnos, pero un modelo de machine learning no puede. Por eso gracias al feature engineering podemos crear datasets mas rico en características para ser explotadas por los modelos.

También como descomprimimos unas variables que estaban comprimidas en una sola para poder formar varias. Como 'deposit_type' transformandose en 'reservo_sin_reembolso' y 'reservo_sin_depositar'.

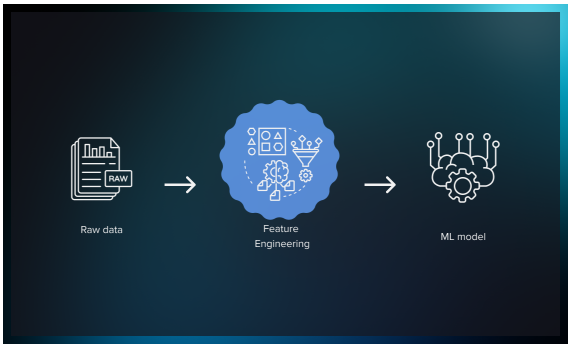

Figura 2: feature engineering

3. Modelos

Creemos que fue muy útil haber entrenado como el primer modelo al árbol de decisión, ya que es uno que nos permite averiguar cuales son las features con mayor importancia (al menos para decision trees), e indagar un poco mas a fondo el dataset. Además terminan siendo las mismas variables que luego sirven para modelos basados en arboles como XGBoost o Random Forest.

Consideramos que estos modelos basados en arboles fueron los mejores haciendo predicciones para este dataset. Creemos que esto puede deberse a la diversidad de features en el dataset, algunas numéricas continuas y no continuas, otras categóricas con one hot encoding, las cuales pueden ser perfectamente aprovechadas por los arboles de decisión.

En cambio otros modelos como SVM o KNN que precisan que los inputs sean estandarizados, no tuvieron muy buenos resultados. Además de que tardaban lo mismo que los otros y daban mucho peores resultados. Consideramos que este es un aspecto que nos quedo pendiente para explorar, principalmente para KNN, donde creemos que realizando un preprocesamiento mejor se podrían haber obtenido mejores resultados.

En cuanto a las redes neuronales, creemos que estas son muy poderosas para muchos tipos de problemas, pero para este caso no son la mejor opción. Concluimos que hay muchas cosas que se pueden cambiar y plantear de maneras diferentes, debido a su gran versatilidad, pero que no pueden ser aprovechadas al máximo para este dataset.

4. Conclusiones

En conclusión creemos que fue un gran trabajo para poder introducirnos en la ciencia de datos y machine learning. Pudimos aprender en la practica como tratar un dataset, explorarlo y utilizarlo para realizar predicciones. Creemos que el hecho de que haya sido con un dataset como reservaciones de hotel, es lo que también haya hecho que sea muy realista y que nos haya gustado tanto.

5. Referencias

- Paper "Hotel booking demand datasets"proporcionado por la cátedra.