# Informe Final - Grupo 28

## Introducción

En este trabajo practico, se realizó un estudio, análisis y entrenamiento de modelos que permitieron llevar a cabo la predicción del valor de una variable, en nuestro caso se trató sobre la cancelación de la reserva.

El trabajo se dividió en cuatro partes:

I) Primera Parte: Análisis Exploratorio y preprocesamiento de los datos del dataset provisto.
II) Segunda Parte: Arboles de Decisión
III) Tercera Parte: Ensambles
IV) Cuarta Parte: Redes Neuronales

## Exploración de Datos

En el checkpoint 1, se realizó un análisis exhaustivo sobre cada variable que se encuentra en el dataset y su influencia con la variable a predecir (en nuestro caso es is_canceled).

Datos Faltantes

Se encontraron en el dataset varias variables con valores nulos:

- Country: Esta variable presenta un porcentaje de 0.3569% de valores nulos. En este caso, decidimos reemplazar los valores nulos por su moda. En este caso es Portugal(PRT).
- Children: Solo hay 4 valores nulos en todo el dataset de esta variable. Tomamos la decisión de reemplazar esos datos nulos con 0.
- Agent: En este caso, existe un 12% de valores nulos en esta variable. En este caso, decidimos reemplazar los valores nulos por 0.
- Company: Esta variable tiene un alto porcentaje de valores nulos (95%), por esta razón decidimos eliminar esta columna del dataset.

Transformaciones

Se realizaron distintas transformaciones para poder entrenar los modelos:

- OneHotEncoding: Se transformo las variables categóricas en variables numéricas mediante la función de panda de dummies.
- StandardScaler: Se llevo todas las variables numéricas no binarias a una misma escala y así tener mejor resultados en nuestros modelos.

Aparte se realizó transformaciones en distintas variables para poder realizar los entrenamientos:

- Reserved_room_type: Se transformo las letras de las habitaciones en números.
- Arrival_date_month: Se transformo los nombres de los meses en su equivalente en números.
- Meal: Según el paper, la variable tenía cuatro diferentes categorías y había una de ellas que podía tomar dos nombres diferentes (undefined y SC). Se transformo en una sola categoría y todos los undefined se transformaron en SC.
- Country: Esta variable tenía muchos países diferentes pero el país Portugal era el más frecuente de ellos. Nosotros hicimos una transformación de ordenar los países por su frecuencia y tomamos los 20 países mas frecuentes y el resto de los países los catalogamos como 'other'.
- Previous_booking_not_canceled: Esta variable fue eliminada del dataset ya que no tenía una correlación fuerte con nuestro target.
- Assigned_room_type: Esta variable se elimina ya que que proporcionaba la misma información que reserved_room_type y lo encontramos redundante para nuestro entrenamiento.

## Modelos

- Árbol de decisión

Se realizo un entrenamiento optimizando los hiperparámetros utilizando el modelo de Random Search Cross Validation. Los parámetros que se utilizaron: ·

- Folds: 10
- Splitter: best
- Min_samples_split: 16
- Min_samples_leaf: 16
- Max_depth: 21

- Criterio: entropy
- Class_weight: balanced

Nuestra métrica F1-Score: 0.8535

- Modelo KNN

Hiperparámetros optimizados para KNN:

- Weights: distance
- N_neighbors : 9
- Metric: Euclidean
- Algorithm: kd_tree

F1 Score: 0.8390

- Modelo SVM

Hiperparámetros optimizados para SVN:

- C: 50
- Gamma: 0.01
- Kernel: rbf

F1 Score: 0.8464

- Modelo Random Forest

Hiperparámetros optimizados para RF:

- criterion= ‘entropy’
- min_samples_leaf=1
- min_samples_split=5
- n_estimators=90

F1 Score: 0.882385

- Modelo XGBoost

Hiperparámetros optimizados para XGBoost:

- max_depth=15
- learning_rate=0.1
- n_estimators=100
- subsambple=0.1
- colsample_bytree=0.8

- gamma=0.1
- objective="binary:logistic"
- random_state=42

F1 Score: 0.88

## Cuadro de Resultados

| Modelo | CHPN | F1-Test | Presicion Test | Recall Test | Accuracy | Kaggle |
|---|---|---|---|---|---|---|
| Arbol de Decision | 2 | 0.8535 | 0.8508 | 0.8561 | 0.8469 | 0.8445 |
| Random Forest | 3 | 0.882385 | 0.887580537 | 0.8772507 | 0.880952 | 0.87611 |
| XGBoosting | 3 | 0.88 | 0.88 | 0.88 | 0.88 | 0.86796 |
| Stacking | 3 | 0.885583 | 0.882301656 | 0.888888 | 0.883076 | 0.87499 |
| Voting | 3 | 0.884365 | 0.886218302 | 0.882520 | 0.8825173 | 0.87487 |
| Red Neuronal | 4 | 0.853513 | 0.852185 | 0.854844 | 0.850184 | 0.84170 |

## Conclusiones generales

En el análisis del trabajo desarrollado, hemos llegado a la conclusión que el mejor modelo para predecir nuestra variable de interés fue RandomForest, en el cual obtuvimos un F1 Score de 0.8823. Este modelo da valores altos con respecto a los otros estudiados y es un buen modelo para llevarlo a la práctica para obtener buenas predicciones de la variable target. Esto sucede porque hicimos un análisis profundo del dataset y analizamos todas sus variables en profundidad y así pudimos ajustar de la manera más correcta sus hiperparámetros y así llegar a un valor alto de métrica.

A medida que exploramos el dataset y analizamos sus variables, pudimos mejorar nuestras métricas. Por ejemplo, en el checkpoint 2 teníamos algunas columnas

eliminadas y por esa razón la métrica nos daba bastante baja. Al volver analizar esas variables los resultados obtenidos mejoraron considerablemente.

Como primer modelo, se analizó el árbol de decisión y tuvo un resultado muy favorable. Es una buena aproximación utilizar este modelo ya que nos proporciona información valiosa sobre el dataset y la métrica es muy buena y es muy rápido de analizar y entrenar. Existe una diferencia entre las métricas de Random Forest (nuestro mejor modelo) y el árbol, pero es una buena aproximación.

## Tareas Realizadas

| Integrante | Promedio Semanal (hs) |
|---|---|
| Cecilia Jurgens | 20 horas semanales |
| Martin Schipani | 10-12 horas semanales |
| Marilyn Nicole Soto | 15 horas semanales |