Informe Final - Grupo 05

Introducción

Este trabajo práctico fue realizado con el objetivo de analizar y predecir las cancelaciones de reservas de hotel. A lo largo del proyecto, realizamos diferentes etapas, las cuales se detallarán a continuación.

Realizamos un análisis exploratorio inicial del dataset y llevamos a cabo el preprocesamiento de los datos. Identificamos las variables presentes en el dataset y las clasificamos en cualitativas y cuantitativas. Realizamos algunas transformaciones en las variables para obtener información nueva y facilitar el trabajo posterior. Además, visualizamos los datos mediante histogramas, barplots y scatterplots para comprender mejor las características del dataset.

Para el CHP2 entrenamos un modelo de predicción utilizando diferentes algoritmos de aprendizaje automático. Para eso primero unificamos el feature engineering en los conjuntos de entrenamiento y test. Utilizamos técnicas como Random Search CV y K-fold Cross Validation para seleccionar los mejores hiper parámetros y evaluar el rendimiento del modelo. Las métricas utilizadas fueron accuracy, recall, precisión y F1-score.

Para el CHP3 realizamos clasificadores utilizando diferentes modelos de aprendizaje automático, como KNN, SVM, Random Forest y XGBoost. Optimizamos los hiper parámetros de cada modelo y seleccionamos los mejores. Luego, ensamblamos los modelos utilizando el método de Voting (hard y soft) y Stacking. Evaluamos las diferentes métricas y determinamos que el stacking demostró ser el más efectivo de estos.

Por último, para el CHP4 desarrollamos una serie de modelos de redes neuronales para clasificación, optimizando sus hiperparámetros y evaluando su rendimiento en datos de entrenamiento. Esto incluyó la búsqueda de los mejores valores de hiperparámetros, como la tasa de aprendizaje, el número de unidades en las capas ocultas y la función de activación. Cada modelo fue evaluado en términos de métricas de clasificación y se visualizó el progreso de su rendimiento en cada uno. Con estas evaluaciones seleccionamos el modelo más adecuado y obtuvimos conclusiones sobre su eficacia en la tarea de clasificación.

Cuadro de Resultados

Modelo	CHPN	F1-Test	Presicion Test	Recall Test	Accuracy	Kaggle
Random Search (Random forest)	2	0.8804043289567 653	0.882003302144 63951	0.8788111427944725	0.87923670050479 84	0.87775
Stacking	3	0,88	0,89	0,89	0,88	0,87
Soft Voting	3	0,87	0,88	0,88	0,87	0,862
modelo_4 (RN)	**4**	**0.901877**	**0.894777**	**0.909090**	**0.899419**	**0.83928**

Random Search (random forest): realizamos una búsqueda de hiper parámetros utilizando Randomized Search CV con criterios como el número de árboles, la profundidad máxima, etc. Luego, entrenamos el modelo con los mejores hiper parámetros y evaluamos su rendimiento en los datos de prueba

Soft Voting: Este modelo considera las probabilidades de predicción de cada variable. Utilizamos los mejores modelos de KNN, Random Forest y XGBoost para crear un clasificador de votación.

Stacking: El modelo Stacking demostró ser el más efectivo, combinando Random Forest, XGBoost y KNN como modelos base y utilizando un Random Forest Classifier como meta-modelo. Tras evaluar el rendimiento de los modelos base mediante validación cruzada, construimos el modelo de Stacking.

Modelo 4: El modelo 4 de redes neuronales demostró ser el más exitoso en nuestra evaluación, superando a los demás en términos de métricas clave como precision, recall, F1 score y accuracy. Este modelo utilizó capas densas en la red neuronal y se benefició de una búsqueda amplia de hiperparámetros que incluyó diversas funciones de activación y ephocs. La combinación de estos hiperparámetros resultó en un rendimiento excepcional, lo que lo convierte en la elección principal para nuestra tarea de clasificación.

Conclusiones generales

El trabajo práctico proporcionó una serie de insights y resultados que son fundamentales para comprender la eficacia de los modelos de clasificación utilizados y el procesamiento de datos previo.

El análisis exploratorio de datos resultó extremadamente útil, ya que permitió una comprensión profunda de la estructura y distribución de las variables en el conjunto de datos. Proporcionó información valiosa sobre las relaciones entre las variables y destacó patrones de correlación positivos y negativos.

Las tareas de preprocesamiento demostraron ser esenciales para limpiar y preparar los datos para su uso en modelos de clasificación. Esto incluyó la eliminación de registros, la recodificación de variables y la gestión de valores faltantes a través de diferentes métodos.

Se exploraron diversos modelos de clasificación, incluyendo K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest y XGBoost. La implementación de ensambles, como el stacking, destacó como una estrategia muy efectiva para mejorar la precisión de las predicciones.

En nuestro caso, las métricas más sobresalientes, como F1, precisión, recall y accuracy se lograron con el Modelo 4 de Redes Neuronales. Este modelo demostró ser una elección sólida para la tarea de clasificación y superó a los demás en términos de rendimiento en los datos de prueba y la competencia en Kaggle.

Para mejorar los resultados, se podrían considerar otras técnicas de procesamiento de datos en futuros trabajos, y la optimización de hiperparámetros para SVM podría ser una dirección a explorar.

Tareas Realizadas

Integrante	Promedio Semanal (hs)
Iara Jolodovsky	5 hs
Martin Abramovich	4 hs
Tomas Vainstein Aranguren	4 hs