Informe entrega nro 2.

Integrantes:

- **Andres Felipe Calvo Ariza CC. 1006616347**
- **Andres Guillermo Toloza Guzman CC.1004843452**

Descripción del Progreso Alcanzado

Este informe documenta nuestro progreso en el ámbito del Aprendizaje Automático, destacando nuestros logros iniciales. Como recién llegados a este campo, hemos tenido la oportunidad de enfrentar desafíos iniciales, entre ellos, los pasos a seguir para desarrollar un modelo y además la limpieza y adaptación de los datos para lograrlo.

Importación de Bibliotecas

Usamos bibliotecas fundamentales de Python, tales como pandas y scikit-learn la cual nos ayudó para la generación de modelo en sus pasos de generación, entrenamiento y puntuación.

Uso de los datos

Para este paso, debimos transformar los datos para cumplir el requisito de tener 5% de datos faltantes en tres columnas, las cuales fueron 'policy_tenure','area_cluster','age_of_car'. Esto se realizó para el archivo de ‘train.csv’.

Una vez logramos adaptar los datos de acuerdo con los requisitos establecidos, procedimos a eliminar la columna ‘policy_id’ la cual nos dimos cuenta no aportaba al modelo dado que es una columna de identificación única como un ID junto a las filas nulas dado que el modelo que vamos a implementar no admite valores nulos y dado que posteriormente vamos a transformar las variables categóricas en números, si hacemos un fillna con 0 por ejemplo, alteraría el resultado y no sería confiable.

Ya teniendo limpio nuestro dataset, empezamos a adaptarlo para poder generar un modelo a partir de él, por lo que debimos transformar los datos de las columnas de variables categóricas en número, aplicando una técnica que descubrimos investigando llamada codificación de etiquetas donde a cada tipo de categoría se le asigna un número. Para esto nos ayudamos del

"LabelEncoder" el cual es proporcionado por scikit-learn. Este proceso lo realizamos para el archivo con el que se generara el modelo (train.csv)

Nota: Para este proceso caímos en cuenta que se debía realizar antes de la generación del modelo pues nos generaba error por el tipo de dato, por lo que nos tocó investigar para poder saber que era necesario este paso.

Modelado

Para nuestro primer modelo, decidimos realizarlo con regresión logística.

Investigando de acuerdo con tipos de modelos de machine learning y sus casos de uso en cuanto a problemas, pudimos llegar a la conclusión que este tipo de modelo nos ayuda con problemas predictivos de una variable objetivo-categórica, que en nuestro caso será si realizó un reclamo o no.

Además, como decidimos que nuestro modelo de machine learning se iba a evaluar mediante el F1 SCORE, esta evaluación es muy recomendada para modelos de clasificación, el cual, regresión logística corresponde a ese tipo.

Lo que realizamos fue usar scikit-learn e inicializar un nuevo modelo de regresión logística, luego dividimos nuestros datos en 30% destinado para testing del modelo y el 70% restante para entrenamiento. Para cada una nuestras variables (training y testing) separamos en una serie de pandas la columna objetivo-llamada 'is_claim' y la data sin esa misma columna.

Posterior a eso, usamos la data sin la columna 'is_claim' y la serie de la misma columna objetivo del archivo train y las usamos para ajustar o entrenar el modelo de regresión logística.

Predicción y precisión

Para este paso, usamos la data que guardamos sin la columna objetivo (la que usamos para hacer predicciones) y la usamos para hacer una predicción usando nuestro modelo que creamos y ajustamos anteriormente.

Al momento de determinar una primera precisión de nuestro modelo, usamos la función llamada "**accuracy_score"** proporcionado por **scikit-learn** y le pasamos la serie de la columna objetivo, pero del archivo de prueba y le pasamos las predicciones realizadas por el modelo (que también se realizaron de la data asignada a testing), este resultado lo multiplicamos por 100 y con esto obtuvimos una primera información en términos de precisión para nuestro modelo, el cual nos arrojó un primer resultado de 93.4%.

Pasos siguientes:

1) Probar un modelo diferente, tenemos planeado probar con un modelo de Random Forest dado que también vimos que es bueno para modelos de predicción de una variable categórica y, además, soporta valores nulos.

2) Implementar el análisis de precisión utilizando el F1 SCORE.

3) Implementar gráfica para entendimiento de resultados.