Proyecto Claims Severity Prediction
ENTREGA 1

INTEGRANTES:

YOHEL OSVALDO PEREZ GARCIA

TATIANA ELIZABETH SÁNCHEZ SANIN

DANIELA ANDREA PAVAS BEDOYA

MATERIA:

Introducción a la Inteligencia Artificial para las Ciencias e Ingenierías

Profesor:

Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

1. Planteamiento del Problema

Para abordar este problema, una compañía de seguros puede recopilar y analizar datos sobre reclamos pasados para identificar patrones o tendencias que puedan ayudar a predecir la gravedad de reclamos futuros. Esto puede implicar analizar una variedad de variables, que incluyen datos demográficos, información de la póliza y detalles sobre la naturaleza del reclamo.

Al usar algoritmos de aprendizaje automático para analizar estas variables, las compañías de seguros pueden construir modelos que puedan predecir con precisión la gravedad de los reclamos de seguros. Esto puede ayudar a las compañías de seguros a evaluar mejor el riesgo asociado con diferentes pólizas y tomar decisiones más informadas sobre cómo gestionar los reclamos. En última instancia, esto puede ayudar a las compañías de seguros a proporcionar servicios más eficaces y eficientes a sus clientes.

2. Datos

Los datos provienen de una competencia de Kaggle, (https://www.kaggle.com/competitions/allstate-claims-severity/data), que contiene datos simulados de reclamos de seguros y detalles de los clientes. El conjunto de datos contiene 188,318 registros y 132 características, incluidas las características categóricas y numéricas.

El conjunto de datos de entrenamiento consta de 188,318 observaciones y 131 variables que incluyen 72 variables categóricas binarias, 43 variables no binarias con 3-326 niveles, 14 variables continuas y la variable de resultado, "pérdida". Dado que todas las variables predictoras se anonimizan, no se divulga información específica sobre ellas. No hay valores perdidos. También se observa un conjunto de datos de prueba con 125,546 filas.

3. Métricas

La métrica de desempeño de machine learning utilizada será el Mean Absolute Error (MAE), ya que es la métrica utilizada en la competencia.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Donde n es la cantidad de datos, y_i los datos observados y $\hat{y}_i$ Los valores pronosticados.

Como métrica de negocio, se podría utilizar la reducción del costo de los reclamos para la compañía de seguros.

4. Desempeño

Un desempeño deseable en producción podría ser tener un modelo que pueda predecir el monto de la reclamación con un MAE de menos de 1200 unidades monetarias (la métrica utilizada en la competencia). De esta forma poder hacer un mejor análisis de la gravedad y costo de los reclamos realizados por los clientes y permitir la toma de decisiones orientada a la reducción de estos en al menos un 5% en comparación con el modelo existente o sin modelo alguno.

5. Bibliografía

Allstate Insurance, "Allstate Claims Severity," Kaggle, 2016. [Online]. Available: https://www.kaggle.com/competitions/allstate-claims-severity/overview. [Accessed: 11-Mar-2023].

Prof, C. (2022, 3 enero). 3 Ways to Calculate the Mean Absolute Error (MAE) in R [Examples]. CodingProf.com. https://www.codingprof.com/3-ways-to-calculate-the-mean-absolute-error-mae-in-r-examples/

ENTREGA 1

Esta primera entrega sólo consta de un archivo en el que (1) describas el problema predictivo a resolver, (2) el dataset que vas a utilizar, (3) las métricas de desempeño requeridas (de machine learning y de negocio); y (4) un primer criterio sobre cual sería el desempeño deseable en producción.

Por ejemplo:

1. dadas las caracetrísticas de una casa (superficie, localización, etc.) vamos predecir el precio de venta de una casa en el mercado.
2. vamos a usar el dataset de kaggle esta competición (poner el enlace), que tiene X número de muestras (casas) y tales columnas (poner la lista de columnas, o si es muy grande, poner las que se consideren más representativas para dar una idea de cómo es el dataset).
3. como métrica de machine learning vamos a usar el MAE (Mean Absolute Error) que es el define la propia competencia (podría ser otro si el de la competencia es complicado). Por ejemplo, si el precio de una casa es 100K y un modelo predice 120K, entonces el error es del 120%. Pero si el precio es de 40K y el modelo predice 20K el error es del 50%. Como métrica de negocio se podría usar el incremento en ventas gracias a la utilización del modelo.
4. si las ventas no aumentan más de un 10% no merece la pena poner el modelo en producción ya que el coste de desarrollo y mantenimiento no cubriría las ganacias adicionales de ese aumento.

tened en cuenta que el proyecto va hasta la métrica de machine learning (tenéis que reportar el rendimiento de los modelos). La métrica de negocio y el punto (4) es más una reflexión de cómo se usaría el modelo es un caso hipotético de que se integrase en la operación de una empresa u organización

Otros ejemplos de los puntos 3 y 4:

- **Ejemplo 1**: nuestro modelo de predicción de la patología X en pacientes debería de tener un porcentaje de acierto >80%, pero también un false negative rate <5%, ya que es una patología grave y es preferible no fallar una detección de un paciente que verdaderamente tiene la patología, aunque eso implique que aumente el número de falsos positivos.
- **Ejemplo 2**: según el departamento de marketing de cierta empresa, un modelo de predicción del siguiente producto que compre un cliente debería de tener un porcentaje de acierto de al menos 50%, ya que se usará el modelo para sugerir

recomendaciones a los usuarios. Si el porcentaje de acierto es menor sería contraproducente porque perderíamos ventas.

Como en cualquier proyecto de analítica, esto supone un **primer** criterio, que probablemente se refine o modifique según se avanza en el proyecto, se entiende mejor el posible desempeño de los modelos, con el cliente se va definiendo cómo se usan los modelos en producción/operación, etc.