Allianz

Prueba para Profesional
Senior de Ciencia de Datos

Juan Carlos Baez Lizarazo

Repositorio git

https://github.com/juanbaezl/Car-Insurance-Claim.git

Agenda / Contenido

1. Entendimiento del Negocio
2. Entendimiento de los Datos
3. Preparación de los Datos
4. Modelado
5. Evaluación del modelo
6. Próximos pasos

Agenda / Contenido

1. **Entendimiento del Negocio**
2. Entendimiento de los Datos
3. Preparación de los Datos
4. Modelado
5. Evaluación del modelo
6. Próximos pasos

Entendimiento del Negocio

Objetivo principal

Desarrollar un modelo de **Scoring de Riesgo** que estime la probabilidad de que un asegurado realice un reclamo

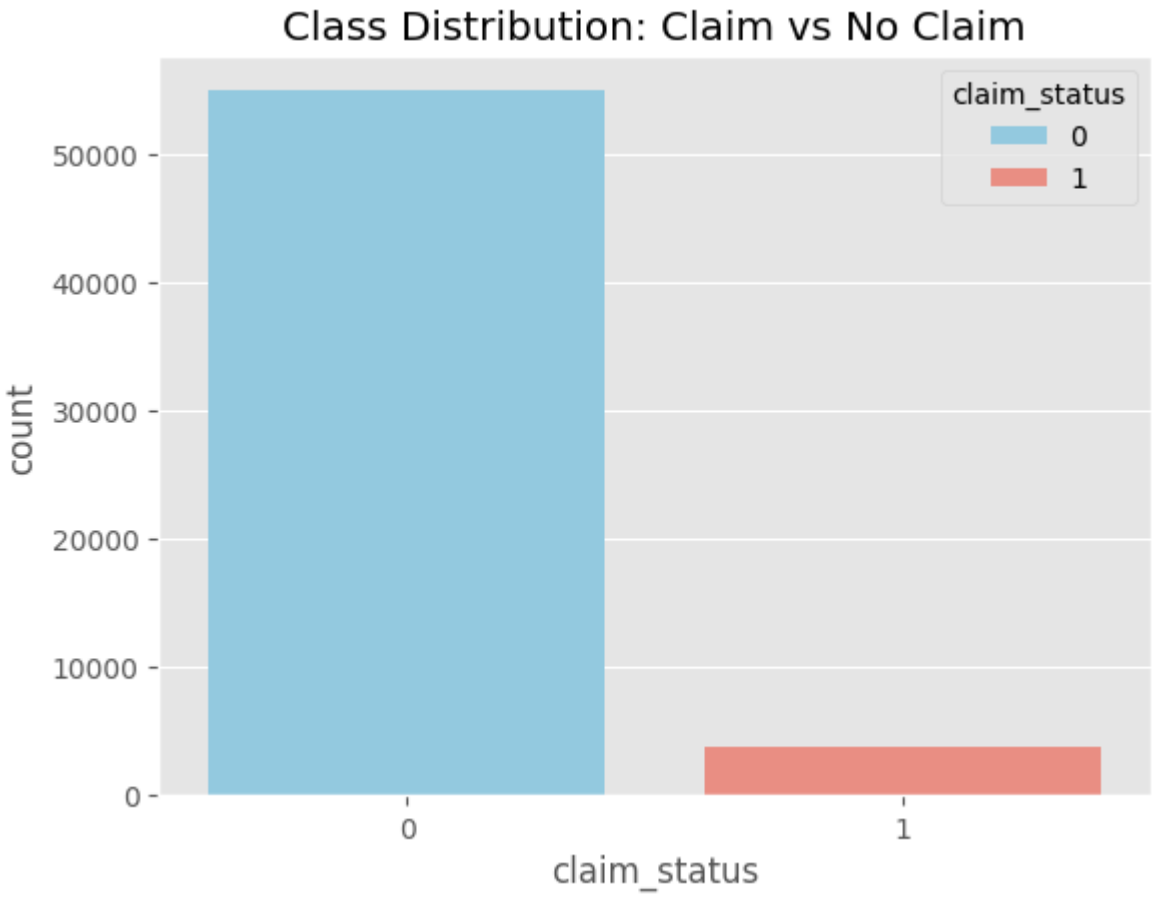
Optimizar tasas para cobrar primas más justas basadas en el riesgo real.

Mejorar el perfilamiento para **Identificar perfiles de alto riesgo** antes de emitir la póliza.

Entendimiento del Negocio

Comportamiento

El modelo de negocio presenta una **asimetría natural** (Desbalance de clases), donde los reclamos constituyen eventos de baja frecuencia pero alto impacto, para garantizar la viabilidad financiera y la precisión analítica, es crítico que el modelo **replique y prediga esta distribución** en sus predicciones



Entendimiento del Negocio

Definición de KPI’s

KPI	Métrica Técnica	Meta Mínima (MVP)	Justificación de Negocio
Capacidad de Discriminación	ROC-AUC Score	**> 0.65**	Garantiza la fiabilidad aceptable para crear *Tiers* (niveles) de precios diferenciados. Un valor inferior implica un alto riesgo de tarifación errónea (sub-cobro o sobre-cobro).
Cobertura de Riesgo	Recall (Sensibilidad)	**> 60%**	Asegura la detección de la mayoría de los reclamos para ajustar correctamente las reservas técnicas, manteniendo un margen tolerable de Falsos Positivos.
Eficiencia de Captación	Lift @ Top 10%	**> 2.0x**	Valida el poder predictivo en el segmento crítico: el grupo clasificado como "Más Riesgoso" debe tener una densidad de reclamos 2 veces mayor al promedio de la cartera.

Agenda / Contenido

1 Entendimiento del Negocio

2 **Entendimiento de los Datos**

3 Preparación de los Datos

4 Modelado

5 Evaluación del modelo

6 Próximos pasos

Entendimiento de los Datos

Estrategia

Auditoría Técnica: Realizaremos un diagnóstico inicial para identificar valores nulos, inconsistencias en tipos de datos (ej. variables numéricas codificadas como texto) y cardinalidad.

Saneamiento Estructural: Antes de cualquier análisis, corregiremos los formatos de las variables para hacerlas legibles por el sistema.

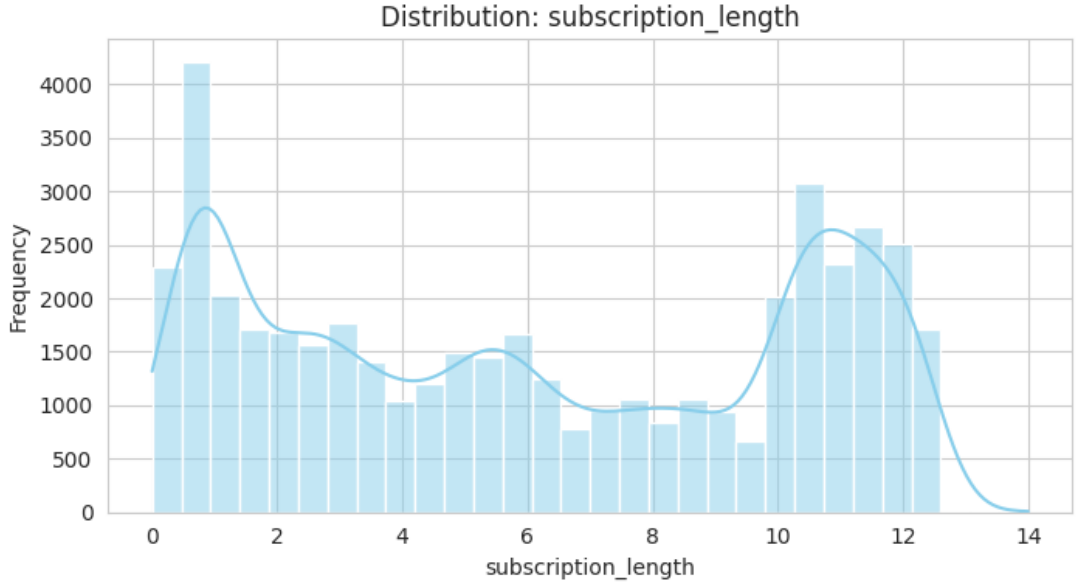
Estrategia de Particionamiento: División de los datos en conjuntos de Entrenamiento siguiendo una validación cruzada estratificada y conjunto de Prueba.

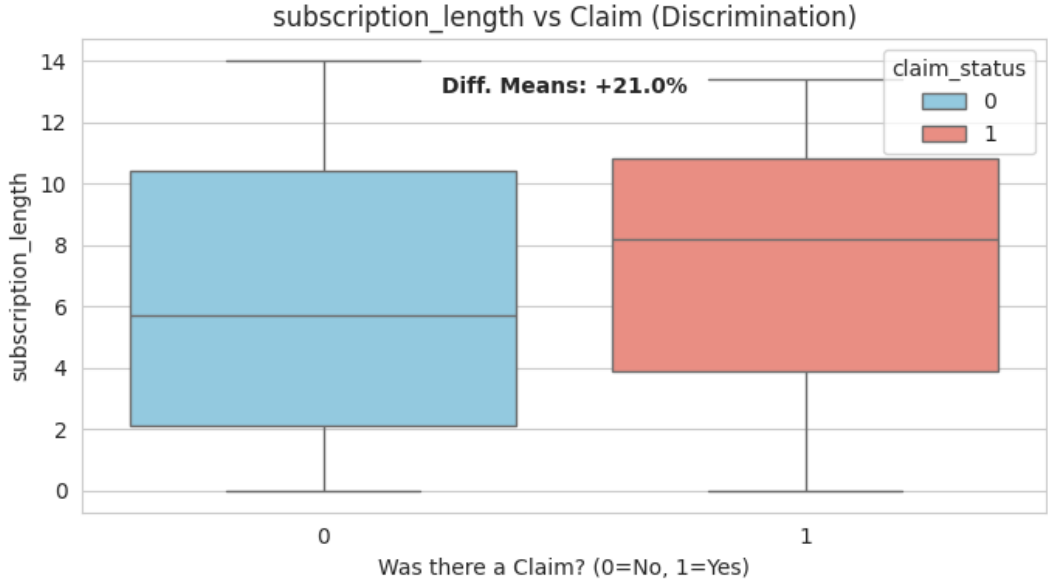
Análisis Exploratorio Profundo: Buscar patrones, correlaciones y drivers de riesgo que expliquen el comportamiento de los reclamos.

Entendimiento de los Datos

Datos relevantes

A **mayor antigüedad** de la póliza es más probable a que realice un **reclamo**.



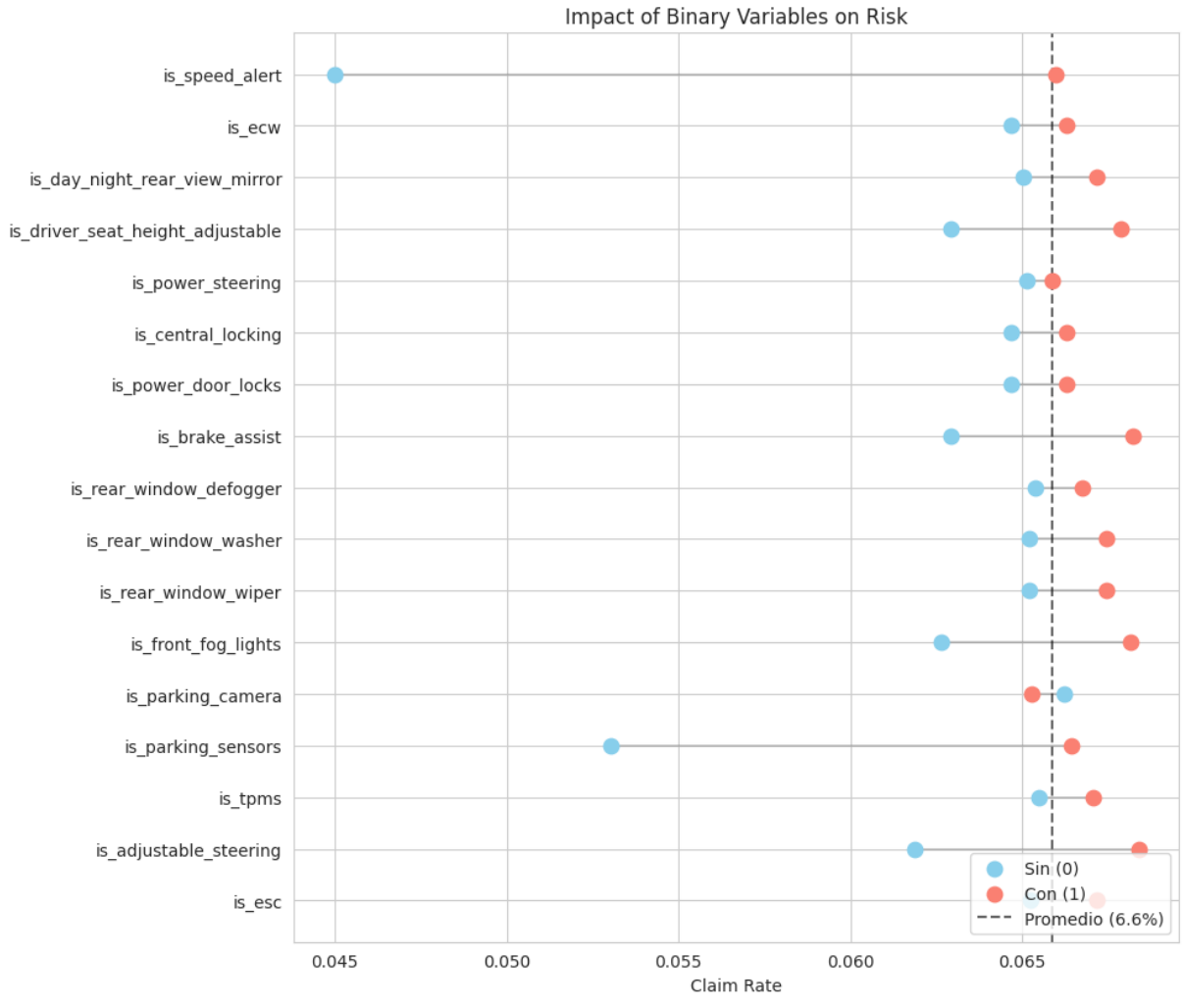


Entendimiento de los Datos

Datos relevantes

Cuando se tiene **equipamiento** es mas probable que realice un reclamo, sobre todo en zonas de choque (ej. Exploradoras)

Esto puede ser explicado por la **gama de valor del carro** y por lo tanto el precio del arreglo (un conductor usaría la póliza para evitar gastos altos)



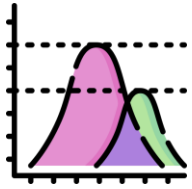
Agenda / Contenido

1 Entendimiento del Negocio
2 Entendimiento de los Datos
3 **Preparación de los Datos**
4 Modelado
5 Evaluación del modelo
6 Próximos pasos

Preparación de los Datos

Estrategia

Reducción de **Redundancia**



Selección Estadística de Características aplicando pruebas estadísticas de dependencia no lineal **(Mutual Information Classification)**

Escalado para evitar darle un peso mayor a alguna característica**.**

Transformación de variables categóricas a numéricas. Se hará uso de **One-Hot Encoding** (nominales) **y Order Encoding** (jerarquía).

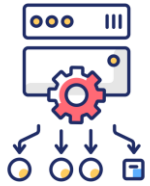
Agenda / Contenido

1 Entendimiento del Negocio
2 Entendimiento de los Datos
3 Preparación de los Datos
4 **Modelado**
5 Evaluación del modelo
6 Próximos pasos

Modelado

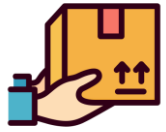
Estrategia

Entrenar 3 modelos **(Random Forest, XGBoost, LightGBM)** siguiendo la validación cruzada, escogiendo el mejor en cuanto al compendio de las 3 métricas objetivo.



Optimización bayesiana para el modelo ganador, aplicando hiperparámetros que reduzcan los reclamos no predichos.

Entrenamiento total con los mejores **parámetros encontrados.**



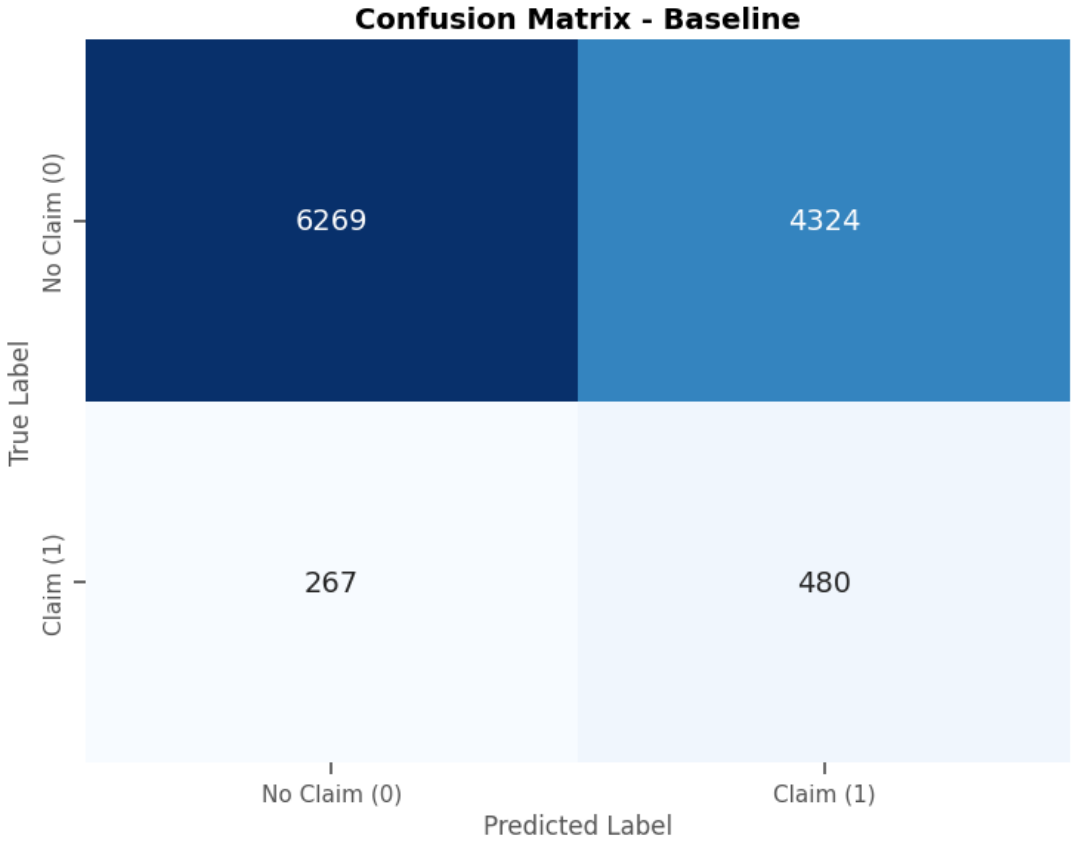
Empaquetado del modelo para la evaluación en el Dataset de test.

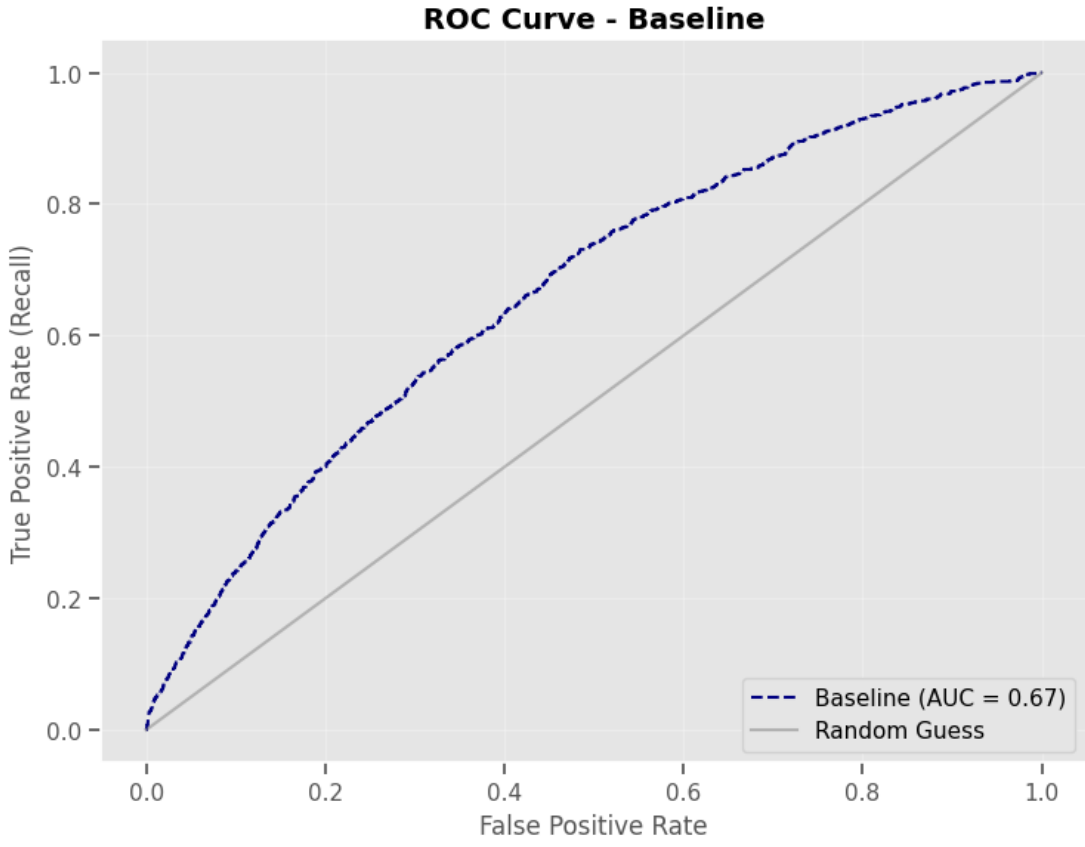
Agenda / Contenido

1 Entendimiento del Negocio
2 Entendimiento de los Datos
3 Preparación de los Datos
4 Modelado
5 **Evaluación del modelo**
6 Próximos pasos

Allianz

Modelado
Resultados





Allianz

Modelado

Resultados

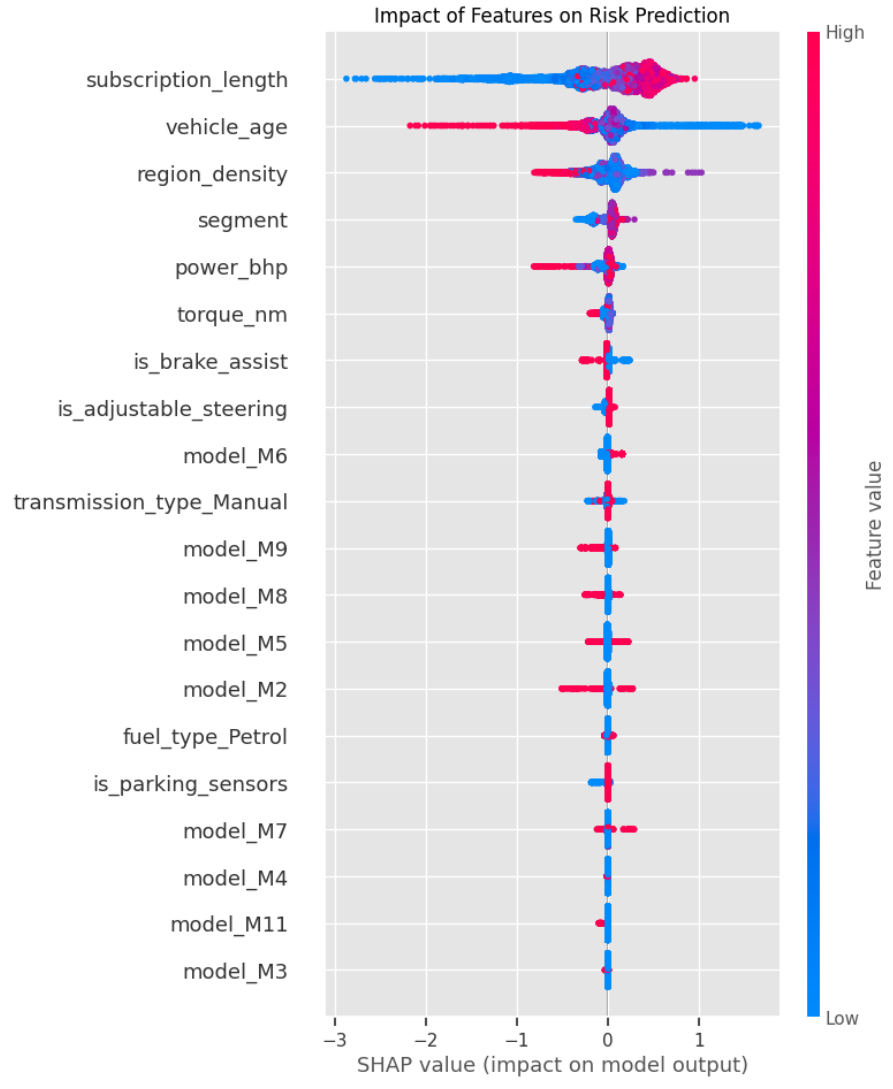
KPI	Métrica Técnica	Meta Mínima (MVP)	Resultado
Capacidad de Discriminación	ROC-AUC Score	**> 0.65**	0.67
Cobertura de Riesgo	Recall (Sensibilidad)	**> 60%**	0.64
Eficiencia de Captación	Lift @ Top 10%	**> 2.0x**	2.26

Modelado

Shap Values

El modelo toma como variables importantes las mencionadas en los datos relevantes donde a **mayor antigüedad** en la póliza mas riesgo de reclamo.

De aquí también podemos ver que entre **mas nuevo el vehículo,** más probabilidad de reclamo y en **ciudades de densidad media** también.



Modelado

Conclusiones

Es posible usar el modelo, pero se necesitaría mas **información del conductor o del vehículo** para hacerlo más robusto.

No siempre la **seguridad alta** en un carro dependa del reclamo, muchas veces eso mismo hace que sea mas posible un reclamo ante un choque leve.

Es recomendado revisar **planes de fidelización y auditoria** de los asegurados.

Agenda / Contenido

1 Entendimiento del Negocio

2 Entendimiento de los Datos

3 Preparación de los Datos

4 Modelado

5 Evaluación del modelo

6 **Próximos pasos**

Modelado

Despliegue de modelo

Es mandatorio el uso de validaciones y despliegues automáticos a Azure **(**en el repositorio se deja instalación de validación de estas con **Github Actions)**

Uso de **Pytest, Bandit, Ruff** para buenas prácticas de codificación **(CI/CD).**

El uso de **Docker** es mandatorio para asegurar **reproducibilidad**

Usar **Data-Store** de **Azure** para el modelo, **Secrets** para variables privadas y pipelines de reentrenamiento con **control de data-drift** teniendo en cuenta la variable principal que es antigüedad de póliza.

Muchas Gracias!