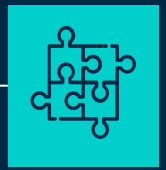
CAR INSURANCE

Comisión: 42390
Profesor: Jorge Ruiz
Tutor:Anderson Ocaña
Alumna: Micaela Valdivia

AGENDA



01

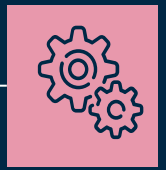
CONTEXTO

Adentrándonos un poco en el mundo de los seguros

02

PREGUNTAS DE INTÉRES

Planteo de preguntas iniciales que vamos a responder con nuestros análisis



03

EDA & INSIGHTS

Análisis Exploratorio de Datos, vamos a examinar y comprender los datos antes de realizar análisis más avanzados o modelado.

04

MACHINE LEARNING

Desarrollo de los Modelos

05

CONCLUSIONES

¿Qué podemos concluir de nuestros modelos?¿Cúal es la decisión final?

CONTEXTO

01

CONTEXTO

En una compañía de seguros de automóviles el precio/ prima que el asegurado debe pagar es fundamental para poder hacer frente al futuro siniestro y los gastos que implica realizar dicho contrato, es por eso que la definición y construcción, de dicho precio, debe ser un buen predictor para garantizar que la compañía pueda hacer frente a futuras contingencias.

En nuestra compañía desafiamos nuestros modelos y scoring, por lo cual atravesamos procesos de cambios:

01

Objetivo del sector de Data Science

Challengear el modelo y poder otorgar al asegurado un precio adecuado según los distintos features y evitar la selección adversa de riesgos con un buen modelo de scoring. De este modo, posicionar a la empresa como gran competidor en el mercado asegurador.

02

Modelos Nuevos:

Vamos a analizar y trabajar con tres diferentes modelos y compararlos para poder definir nuestro mejor estimador:

PREGUNTAS DE INTÉRES

02

Siendo el **objetivo** de poder predecir que ocurra o no un siniestro, y de esta manera ofrecerle un óptimo precio a los asegurados a la hora de ofrecerles una póliza, nos vamos a encontrar con un problema de clasificación. Pero al pensar en nuestro modelo, ¿podemos considerar que entre más factores agreguemos mejor predictor será ?

Nuestras variables

Nos preguntamos ...

01 ¿A mayor cantidad de años de experiencia implica mayor exposición y mayor cantidad de siniestros o menor cantidad de accidentes por mejor desempeño del conductor? Y, que tenga mayor kilometraje de uso ¿implica también mayor cantidad de accidentes? De este modo la compañía podrá alcanzar una selección de los mejores factores de scoreo en su pricing, y no haya selección adversa de riesgos.

02 Siendo una de nuestras variables de interés, el kilometraje, ¿cómo será dicha variable, si analizamos considerando sub segmentos / agrupaciones, según sexo y según ingreso? ejemplo, un hombre de nivel socioeconómico pobre ¿maneja mucho menos en promedio que la mediana, por su bajo nivel de ingreso ?

03 Siendo otra de nuestras variables de interés, el credit score ¿cómo será dicha variable, si analizamos considerando sub segmentos / agrupaciones, según ingreso?

04 ¿Cuánto más multas tiene un asegurado, mayor es la probabilidad que cometa un accidente? O, ¿ Cuánto menos educación o más antiguo sea el auto mayor cantidad de accidentes de tráfico?

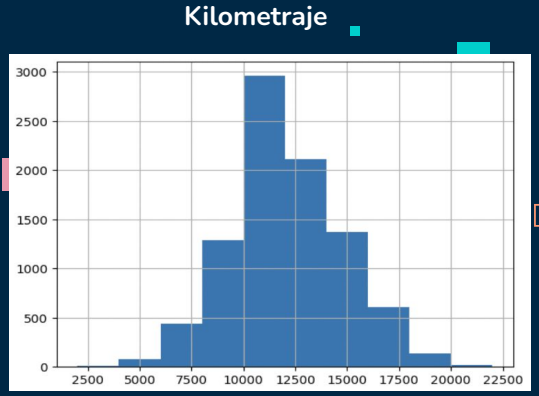
EDA & INSIGHTS

03

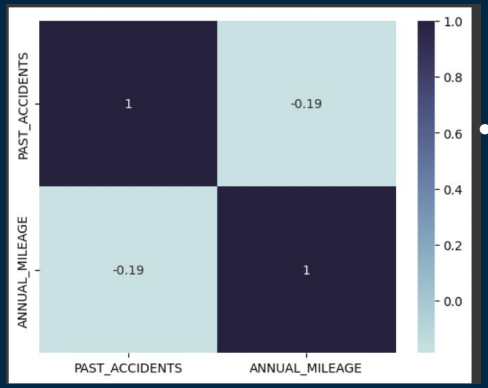
¿Que un conductor tenga más experiencia implica menos siniestros?

Que un usuario tenga más kilometraje de uso en su automóvil, puede significar dos situaciones para una compañía de seguros, que su asegurado es experimentado y tiene menos propensión a sufrir un siniestro, o tiene mayor exposición constante, en por ejemplo salir a la calle diariamente para ir a su trabajo.

- Primero analizaremos la variable **kilometraje**:

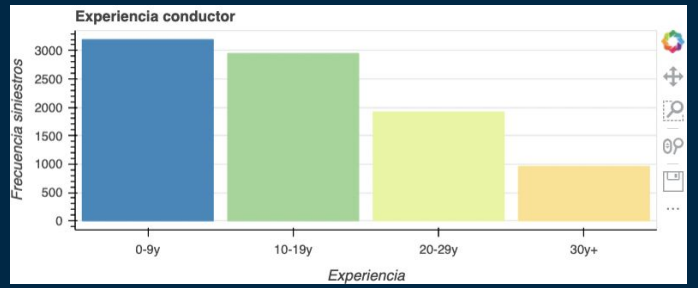


media = mediana =11.693.5
Simétrico



- Segundo veremos como es la correlación entre el **kilometraje y los accidentes de tráfico**:
  INSIGHT: A mayor cantidad de uso que los asegurados le den a su auto, definiendo uso cómo el kilometraje, no vemos una mayor cantidad de accidentes.

- Tercero analizamos la **experiencia del conductor** (medidas en años):
  INSIGHT: La experiencia que tiene un conductor a la hora de manejar su vehículo muestra menor siniestralidad, es decir, estas barras escalonadas, nos indica que a mayor experiencia menor siniestros.

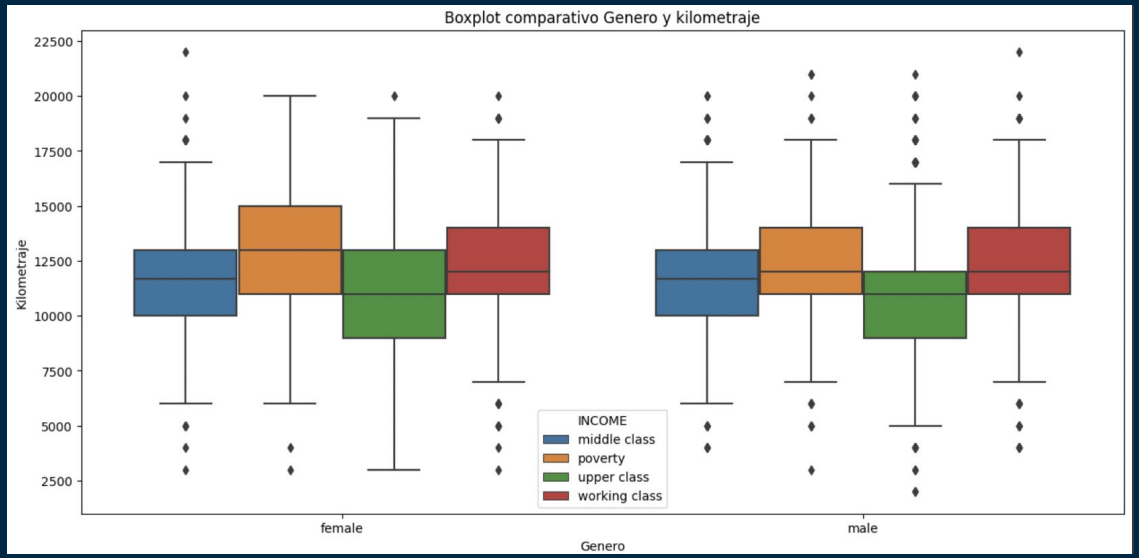


¿Según el ingreso y/o sexo un asegurado usa más o menos su automóvil?

Mediante gráficos de Box Plot:

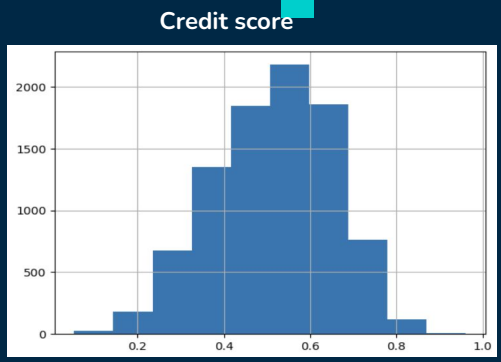
INSIGHT: vemos que casi todos los grupos tienen un kilometraje muy cerca del promedio pero algunos grupos como, hombres de upper class y poverty, y ambos sexos en working class, tienen una asimetría, es decir, su media está alejada de su mediana. Entonces, en base a nuestra pregunta de interés, un hombre de ingreso bajo (poverty) en promedio usa menos que la mediana.

Y por último, hay una cantidad considerable de outliers en segmentos de upper class hombres y middle class mujeres.



¿Que un usuario tenga un mayor ingreso implica que tenga un mejor score crediticio?

- Primero analizaremos la variable **Credit Score**:

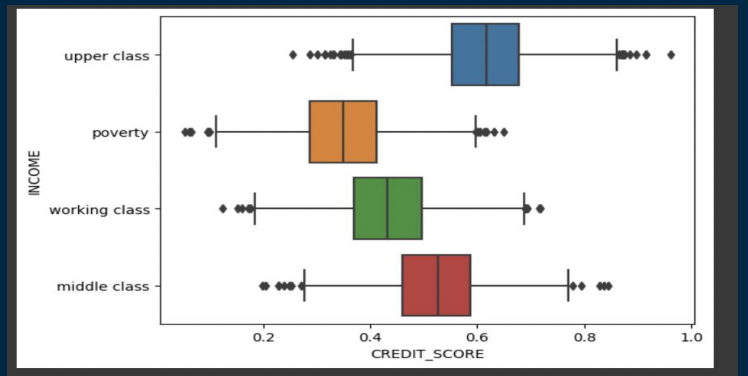


media = mediana =0.5
Simétrico

- Segundo, variable **Credit Score respecto nivel socieconómico**:

- Tercero trabajaremos con el gráfico Box Plot para obtener conclusiones de la variable Credit Score según nivel socioeconómico:

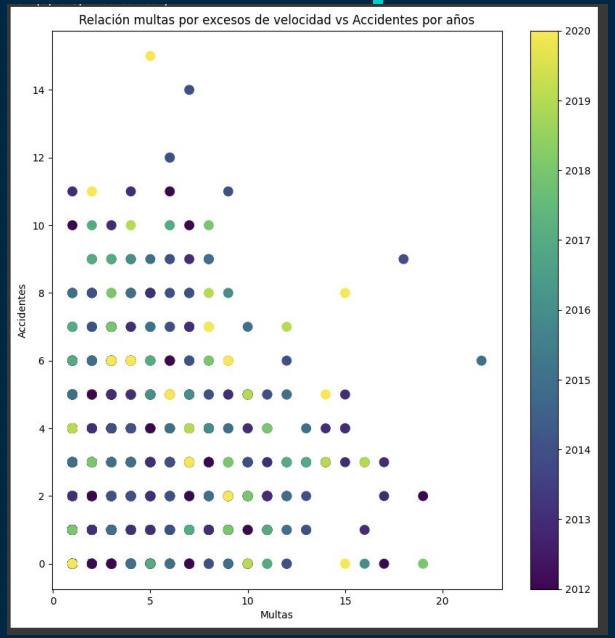
INSIGHT: Podemos ver de manera más clara como la clase "upper" tiene un mayor score. Vemos para todos los casos que la distribución es simétrica

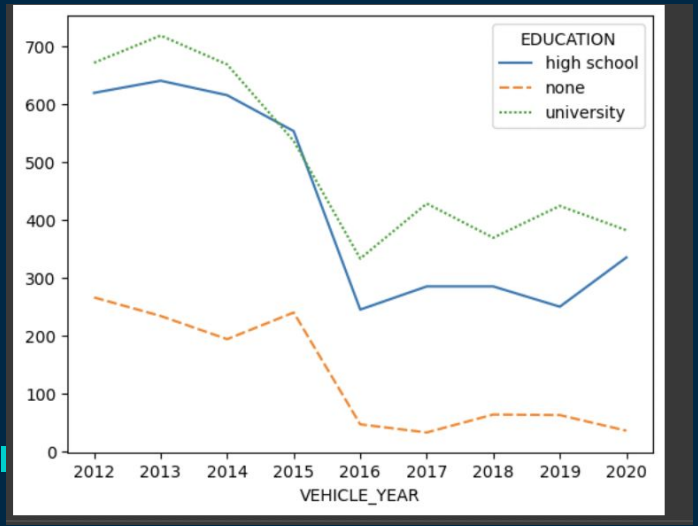


¿Cuánto más multas tiene un asegurado, mayor es la probabilidad que cometa un accidente? O, ¿ Cuánto menos educación o más antiguo sea el auto mayor cantidad de accidentes de tráfico?

- Relación **multas por excesos de velocidad vs Accidentes por años de auto**:

INSIGHT: Vemos una concentración cerca del cero, pero no vemos una relación, una correlación 0, es decir la cantidad de multas por exceso en velocidad no determina que haya más accidentes, en nuestra muestra.





- **Accidentes según el año del vehículo agrupado por educación**:

INSIGHT: Confirmamos que a mayor antigüedad de autos, mayor accidentes, y además podemos concluir que aquellos sin estudio tienen menos accidentes que alguien que tiene un título universitario o secundario.

MACHINE LEARNING 04

NUESTRA VARIABLE TARGET ES DICOTÓMICA

PROBLEMA DE CLASIFICACIÓN

NUESTRO MODELO ACTUAL

REGRESIÓN LOGÍSTICA

PRIMEROS MODELOS PROPUESTOS

REGRESIÓN LOGÍSTICA CON PCA

ÁRBOL DE DECISIÓN

RESULTADOS DE PERFORMANCE DE MODELOS

Modelo	Accuracy	Precisión	Recall	F1 score	Área Curva ROC
Regresión Logística	72.8%	59.9%	35.0%	44.6%	68.3%
Regresión Logística (CON PCA)	82.8%	74.9%	67.6%	71.1%	80.4%
Árbol de Decisión (CON PCA)	83.5%	76.7%	72,3%	74.0%	82.0%

INSIGHT: VEMOS SÓLO CONSIDERANDO LAS MÉTRICAS DE EVALUACIÓN DE LOS MODELOS, EL MODELO DE **ÁRBOL DE DECISIÓN** PODRÍA SER UNA POTENCIAL SELECCIÓN PARA NUESTRO MODELADO.

NUEVO MODELO Y SU OPTIMIZACIÓN

NUEVO MODELO

RANDOM FOREST

OPTIMIZACIÓN

- **Random Forest con Grid Search**
- **Random Forest con Grid Search + RandomOverSampler**
- **Random Forest con Grid Search + RandomOverSampler + MCA**
- **RANDOM FOREST - CROSS VALIDATION**

RESULTADOS DE PERFORMANCE DE MODELOS

Modelo	Accuracy	Precisión	Recall	F1 score
Random Forest con Grid Search	82.4%	75.7%	63,2%	69.3%
Random Forest con Grid Search + RandomOverSampler	83.0%	70.0%	83,0%	76.0%
Random Forest con Grid Search + RandomOverSampler + MCA	83.0%	69.0%	83,0%	76.0%

INSIGHT: NUESTRO ACCURACY CON EL BALANCE DE LA BASE LOGRADO POR EL RANDOM OVERSAMPLER, SI BIEN LA PRECISIÓN DISMINUYE EL RECALL AUMENTA CONSIDERABLEMENTE. POR OTRO LADO, GRACIAS AL GRID SEARCH PUDIMOS ENCONTRAR NUESTROS MEJORES HIPERPARAMETROS DEL MODELO DE RANDOM FOREST

RANDOM FOREST CROSS VALIDATION

TÉCNICAS	SUMMARY	Accuracy
Stratified K - fold		86.4%
K - fold		86.6%

	mean_train_score	mean_test_score	params
2	0.872	0.831	{'max_depth': 10, 'n_estimators': 100}
3	0.872	0.830	{'max_depth': 10, 'n_estimators': 200}
1	0.811	0.807	{'max_depth': 5, 'n_estimators': 200}
0	0.809	0.805	{'max_depth': 5, 'n_estimators': 100}

	mean_train_score	mean_test_score	params
3	0.871	0.828	{'max_depth': 10, 'n_estimators': 200}
2	0.871	0.826	{'max_depth': 10, 'n_estimators': 100}
0	0.809	0.804	{'max_depth': 5, 'n_estimators': 100}
1	0.811	0.804	{'max_depth': 5, 'n_estimators': 200}

INSIGHT: 86.6% vs 86.4%, nosotros estamos prediciendo que ocurra o no un siniestro, lo cual la precisión es importante para ver qué precio cobrar al asegurado, pero en cuanto a cantidad podemos decir que ese 0.2% podría no llega a ser tan representativo, pero optamos por K fold.

CONCLUSIONES DECISIÓN FINAL

05

NUEVO MODELO Y SU OPTIMIZACIÓN

Si comparamos el 87% de accuracy del re entrenamiento con K fold vs el modelo del árbol de decisión (accuracy =86%) y comparamos vs el último modelo que buscamos optimizar hiperparametros y con MCA, donde la accuracy es 83%. HAY MEJORAS EN LA PERFORMANCE DEL MODELO

Con las técnicas de cross validation, se obtiene una evaluación más robusta del rendimiento del modelo al considerar múltiples particiones del conjunto de datos. Esto puede ayudar a tener una mejor estimación del rendimiento del modelo en datos no vistos y reducir el riesgo de sobreajuste, qué es lo que principalmente estamos buscando.

DECISIÓN FINAL - MODELO SELECCIONADO:

Modelo Random Forest considerando la búsqueda de los mejores hiperparametros y con la técnica de Cross Validation (K fold). Ya que no solo obtenemos mejoras en la performance del modelo a nivel métricas sino también un equilibrio entre bias y varianza, buscando el trade-off sesgo-varianza.