CAR INSURANCE CLAIM PREDICTION

By CALISTUS MWONGA

06-06-2025

OUTLINE

I. Business problem

II. Objectives

III. Data overview

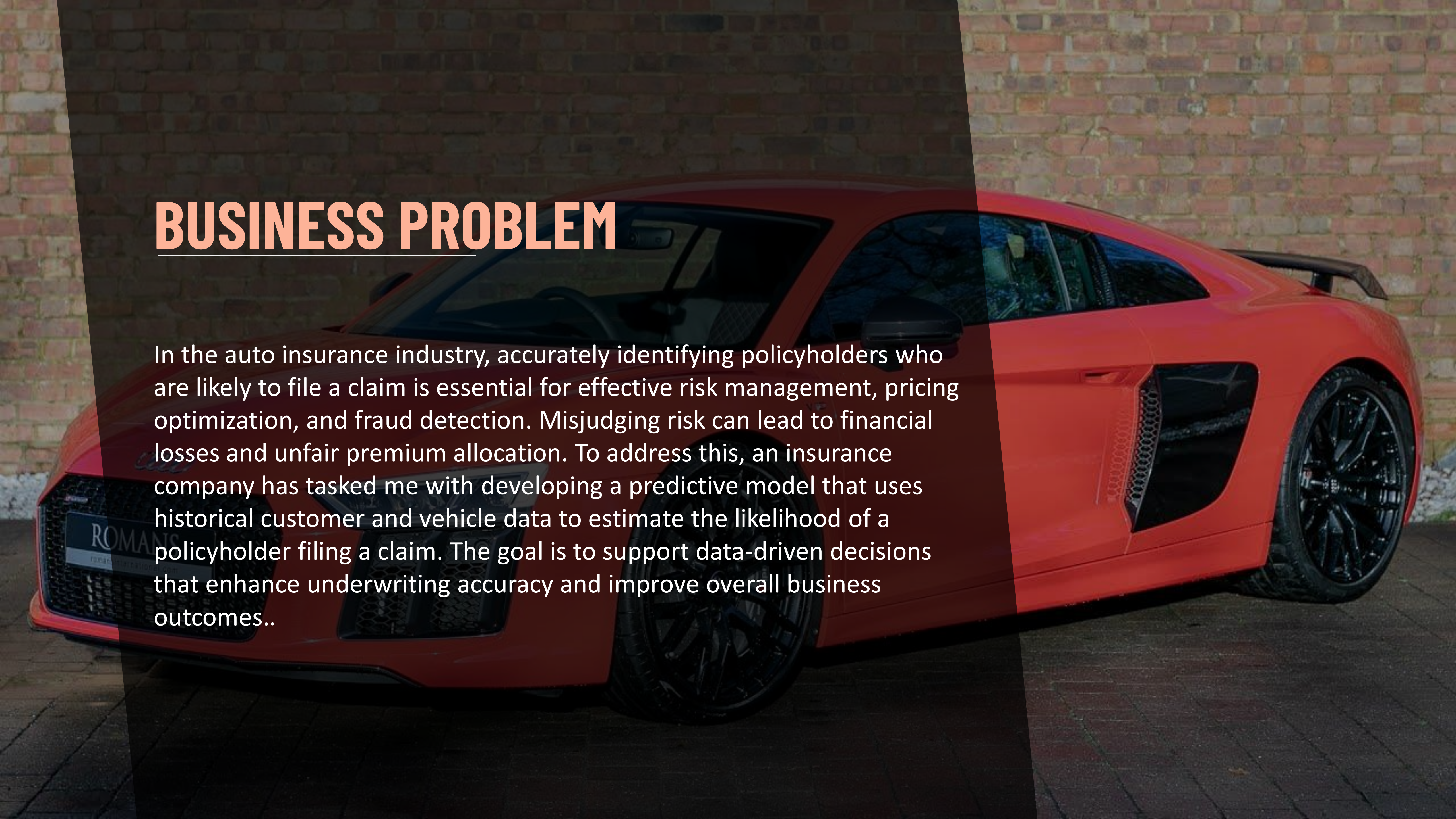
IV. Visualizations and Results

V. Business recommendations

VI. Conclusions

BUSINESS PROBLEM

In the auto insurance industry, accurately identifying policyholders who are likely to file a claim is essential for effective risk management, pricing optimization, and fraud detection. Misjudging risk can lead to financial losses and unfair premium allocation. To address this, an insurance company has tasked me with developing a predictive model that uses historical customer and vehicle data to estimate the likelihood of a policyholder filing a claim. The goal is to support data-driven decisions that enhance underwriting accuracy and improve overall business outcomes..



OBJECTIVES

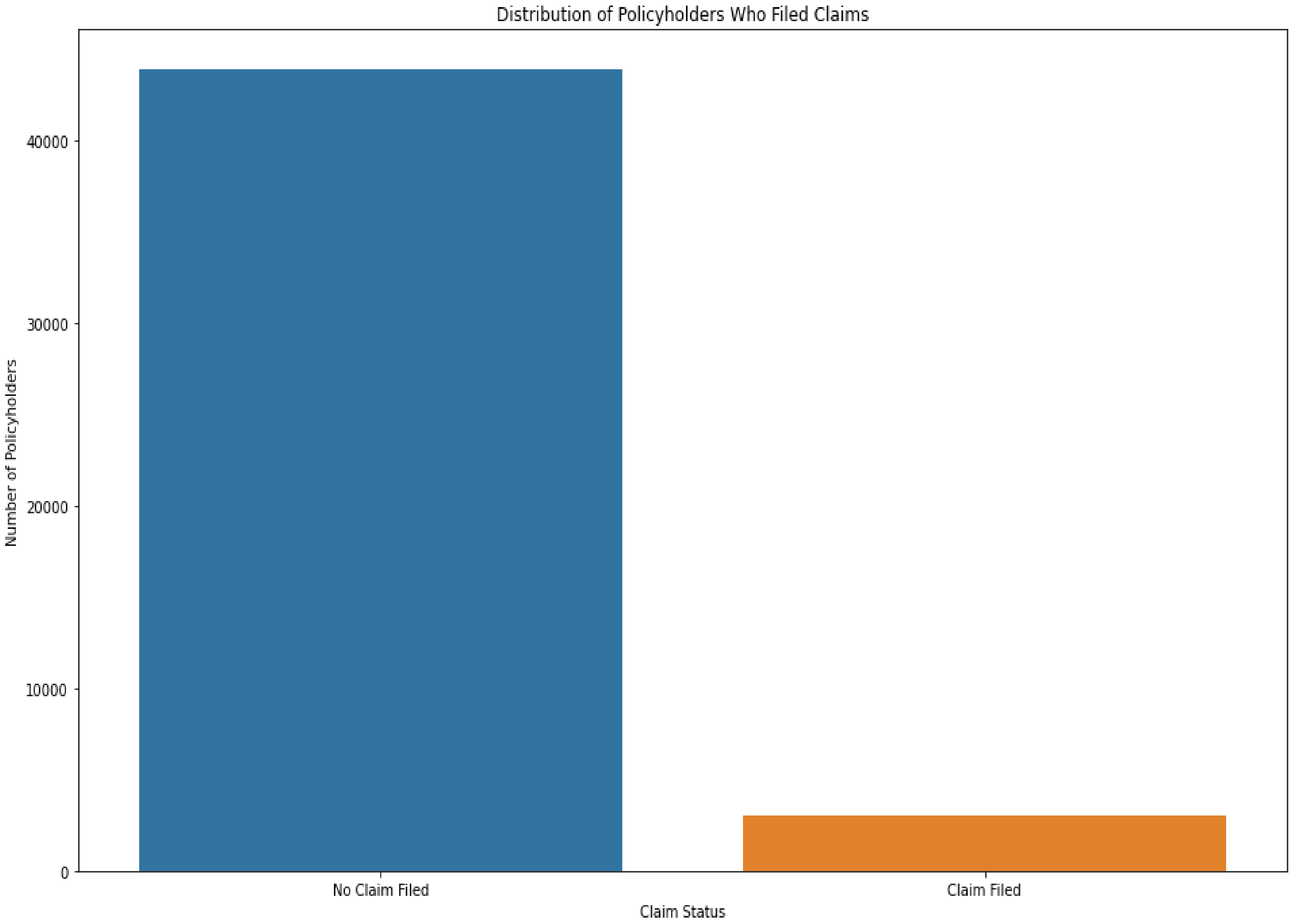
- Explore and Apply various feature engineering techniques to improve model interpretability
- To build classification models that predict the likelihood of a policyholder filing a claim
- To improve the AUC-ROC score by around 10% through hyperparameter tuning of the best-performing models

DATA OVERVIEW

The dataset was obtained from Kaggle. It contains records of car insurance policyholders with various features from policyholder details to vehicle characteristics. The main objective of using this dataset is to predict whether a customer is likely to file a claim. This makes it a classification problem.

Moreover, the dataset contains 44 features and about 58,000 records. The target variable is highly imbalanced with only 6% of the records indicate that a claim was filed. This imbalance posses a challenge for model performance. This necessitates the need to use appropriate evaluation metrics during model development

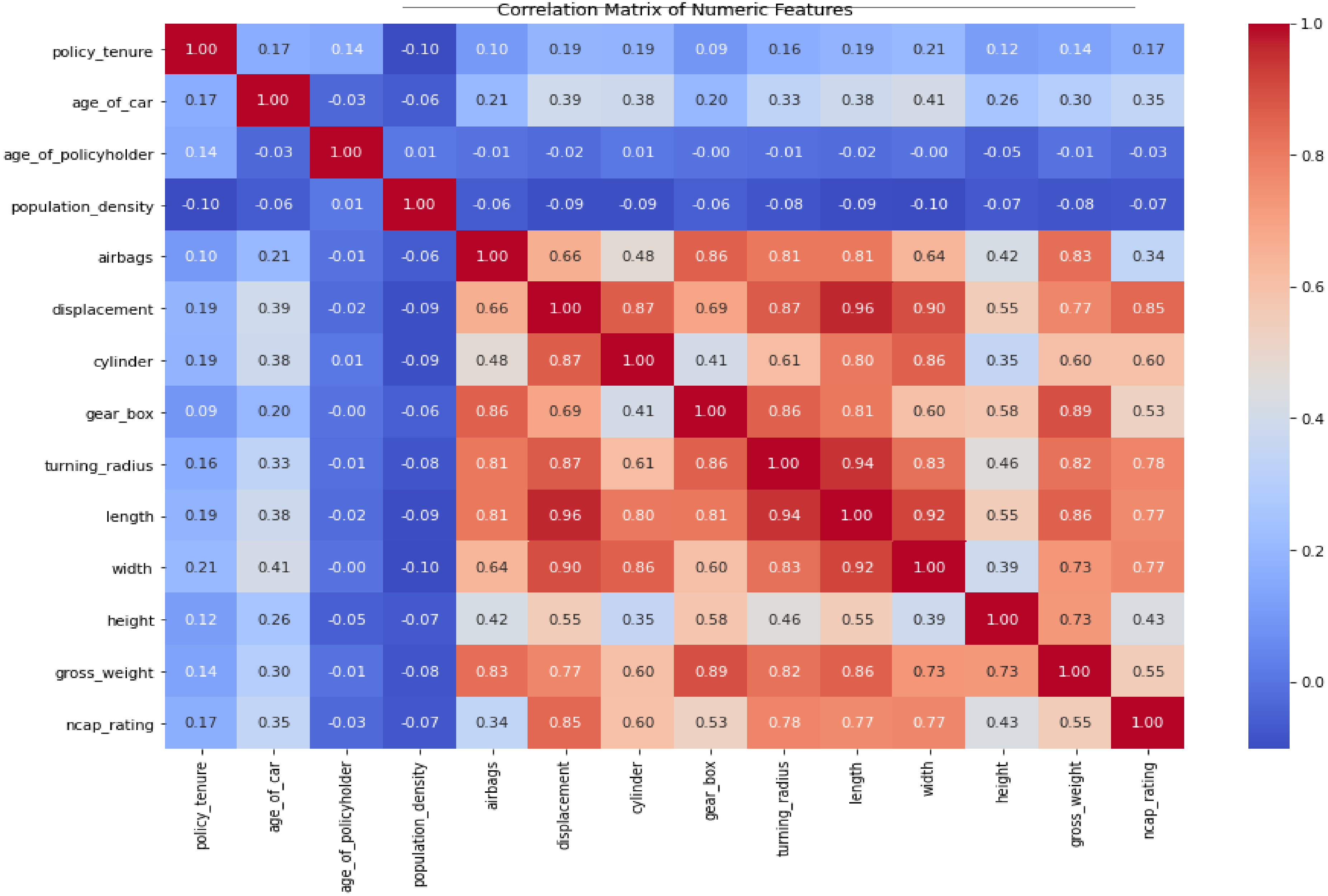
VISUALIZATIONS AND RESULTS



Insight: As you can see, the dataset is heavily imbalanced. The vast majority of policyholders do not file a claim while only a small fraction filed a claim

Percentage wise approximately **6.4%** file claims while **93.6%** do not file claims.

CORRELATION HEATMAP



Correlation Matrix of Numeric Features

	policy_tenure	age_of_car	age_of_policyholder	population_density	airbags	displacement	cylinder	gear_box	turning_radius	length	width	height	gross_weight	ncap_rating
policy_tenure	1.00	0.17	0.14	-0.10	0.10	0.19	0.19	0.09	0.16	0.19	0.21	0.12	0.14	0.17
age_of_car	0.17	1.00	-0.03	-0.06	0.21	0.39	0.38	0.20	0.33	0.38	0.41	0.26	0.30	0.35
age_of_policyholder	0.14	-0.03	1.00	0.01	-0.01	-0.02	0.01	-0.00	-0.01	-0.02	-0.00	-0.05	-0.01	-0.03
population_density	-0.10	-0.06	0.01	1.00	-0.06	-0.09	-0.09	-0.06	-0.08	-0.09	-0.10	-0.07	-0.08	-0.07
airbags	0.10	0.21	-0.01	-0.06	1.00	0.66	0.48	0.86	0.81	0.81	0.64	0.42	0.83	0.34
displacement	0.19	0.39	-0.02	-0.09	0.66	1.00	0.87	0.69	0.87	0.96	0.90	0.55	0.77	0.85
cylinder	0.19	0.38	0.01	-0.09	0.48	0.87	1.00	0.41	0.61	0.80	0.86	0.35	0.60	0.60
gear_box	0.09	0.20	-0.00	-0.06	0.86	0.69	0.41	1.00	0.86	0.81	0.60	0.58	0.89	0.53
turning_radius	0.16	0.33	-0.01	-0.08	0.81	0.87	0.61	0.86	1.00	0.94	0.83	0.46	0.82	0.78
length	0.19	0.38	-0.02	-0.09	0.81	0.96	0.80	0.81	0.94	1.00	0.92	0.55	0.86	0.77
width	0.21	0.41	-0.00	-0.10	0.64	0.90	0.86	0.60	0.83	0.92	1.00	0.39	0.73	0.77
height	0.12	0.26	-0.05	-0.07	0.42	0.55	0.35	0.58	0.46	0.55	0.39	1.00	0.73	0.43
gross_weight	0.14	0.30	-0.01	-0.08	0.83	0.77	0.60	0.89	0.82	0.86	0.73	0.73	1.00	0.55
ncap_rating	0.17	0.35	-0.03	-0.07	0.34	0.85	0.60	0.53	0.78	0.77	0.77	0.43	0.55	1.00

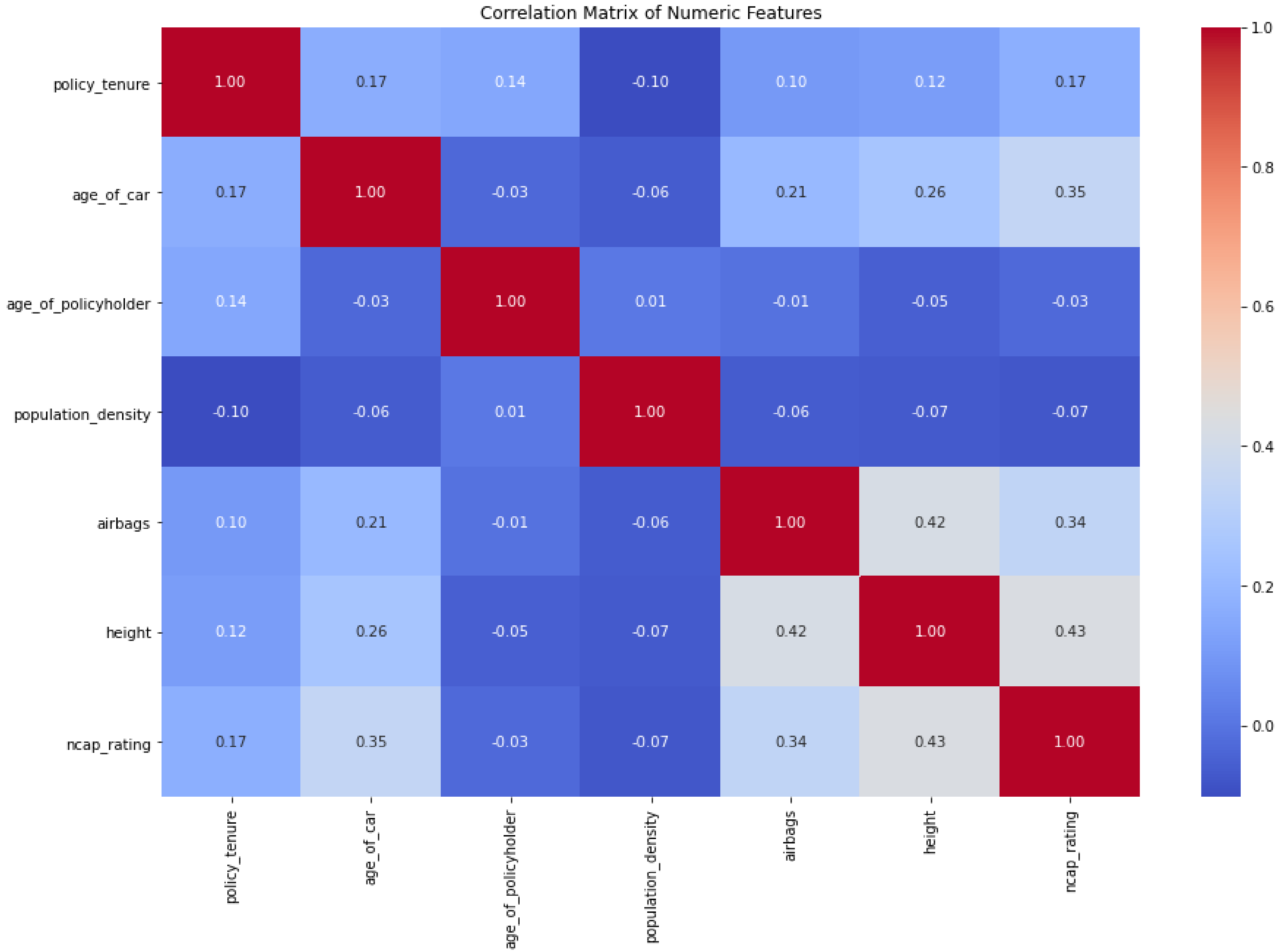
Insight: **F**eatures with correlation coefficient grater than **0.80** are identified as highly correlated.

The decision of which feature to keep was based on which has direct association with the likelihood of a claim being filed.

Among the highly correlated, the features kept were:

- **Airbags** - Directly linked to passenger safety
- **ncap_rating** - This is a standardized safety rating based on crash tests done thus a reliable feature to include in the model

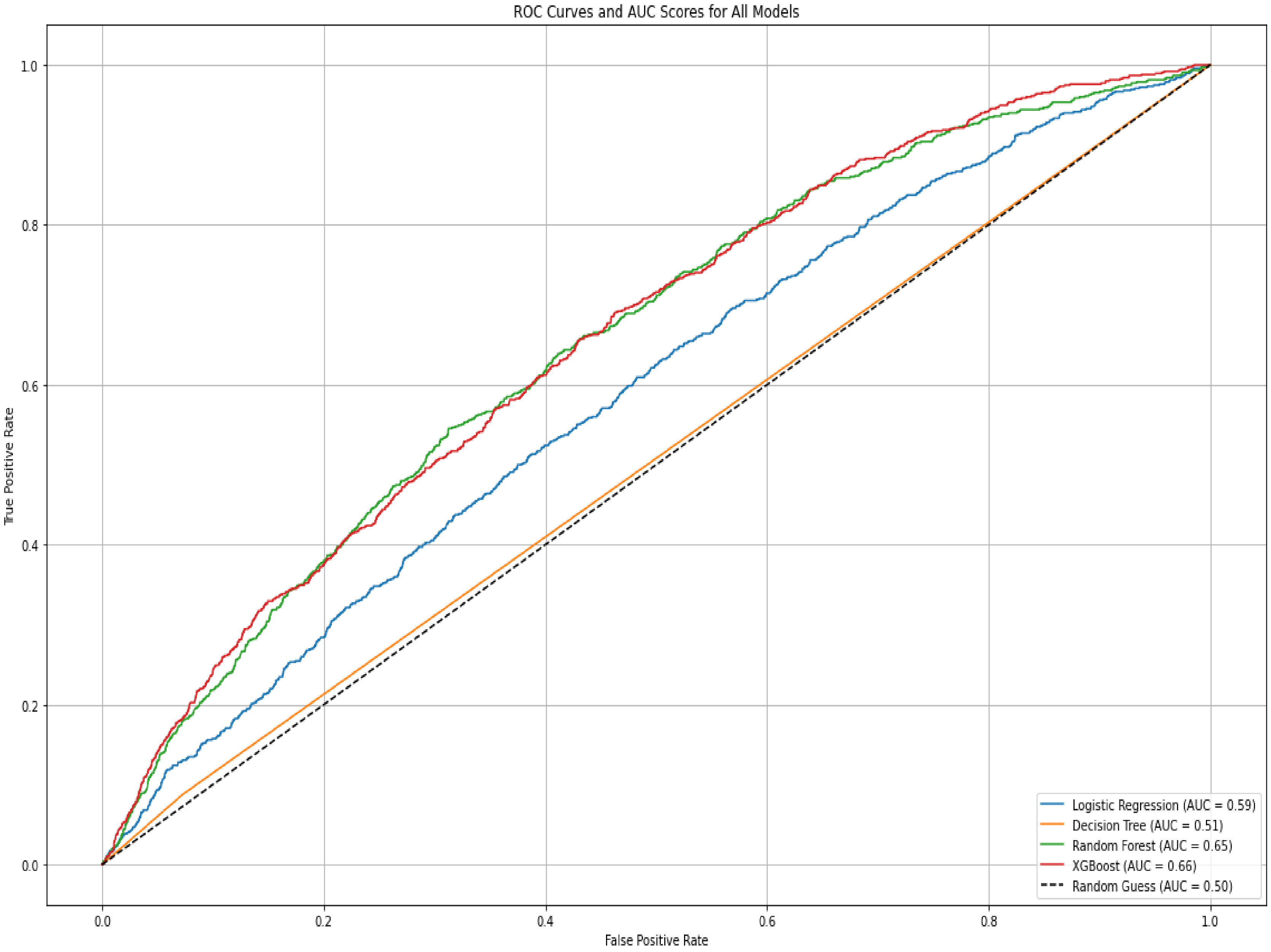
CORRELATION HEATMAP (RESULT)



	policy_tenure	age_of_car	age_of_policyholder	population_density	airbags	height	ncap_rating
policy_tenure	1.00	0.17	0.14	-0.10	0.10	0.12	0.17
age_of_car	0.17	1.00	-0.03	-0.06	0.21	0.26	0.35
age_of_policyholder	0.14	-0.03	1.00	0.01	-0.01	-0.05	-0.03
population_density	-0.10	-0.06	0.01	1.00	-0.06	-0.07	-0.07
airbags	0.10	0.21	-0.01	-0.06	1.00	0.42	0.34
height	0.12	0.26	-0.05	-0.07	0.42	1.00	0.43
ncap_rating	0.17	0.35	-0.03	-0.07	0.34	0.43	1.00

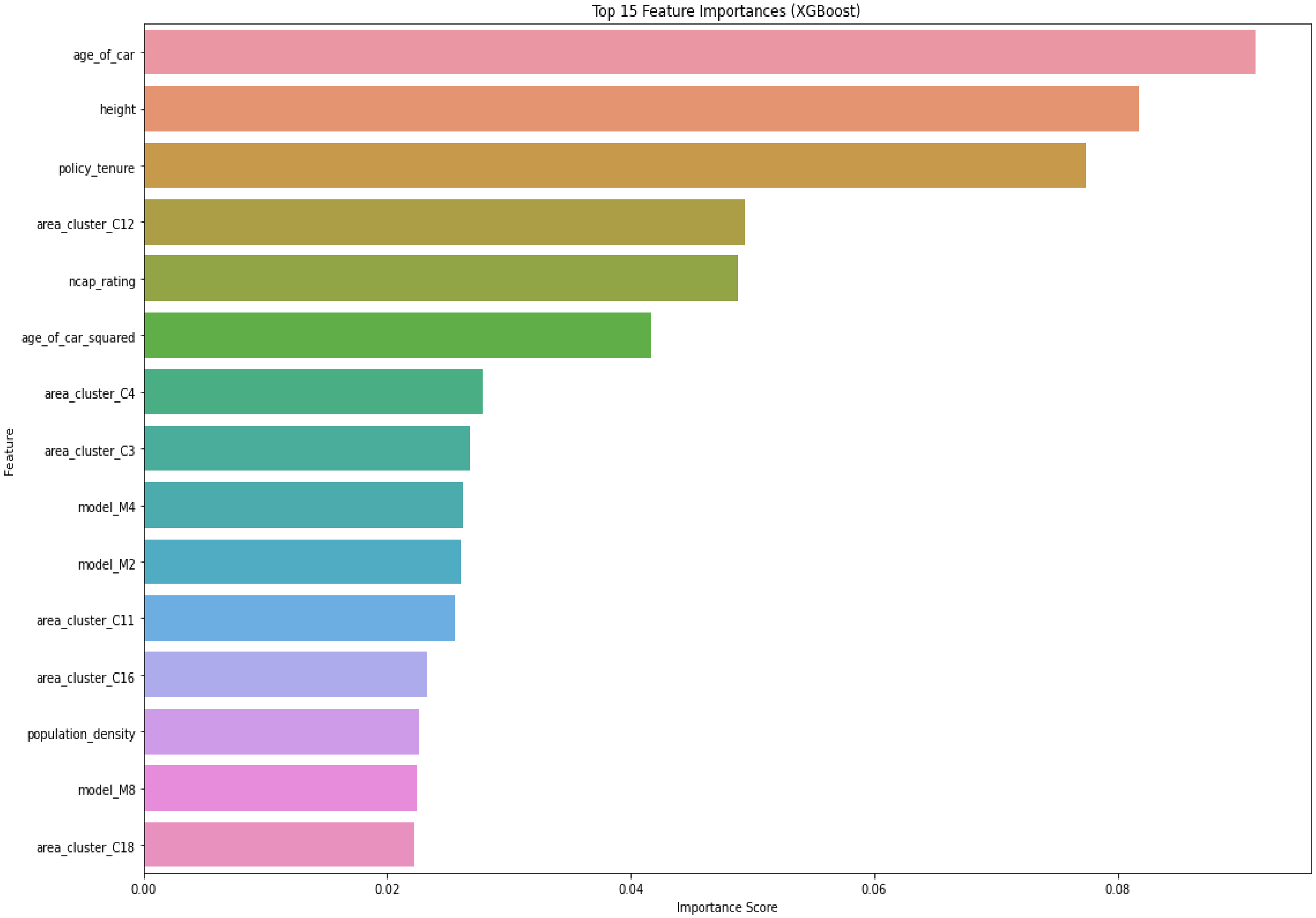
***Insight:** The result after feature selection is fewer number of numeric features where there is no high correlation which can lead to modeling complexity*

ROC CURVES AND AUC SCORES FOR ALL MODELS



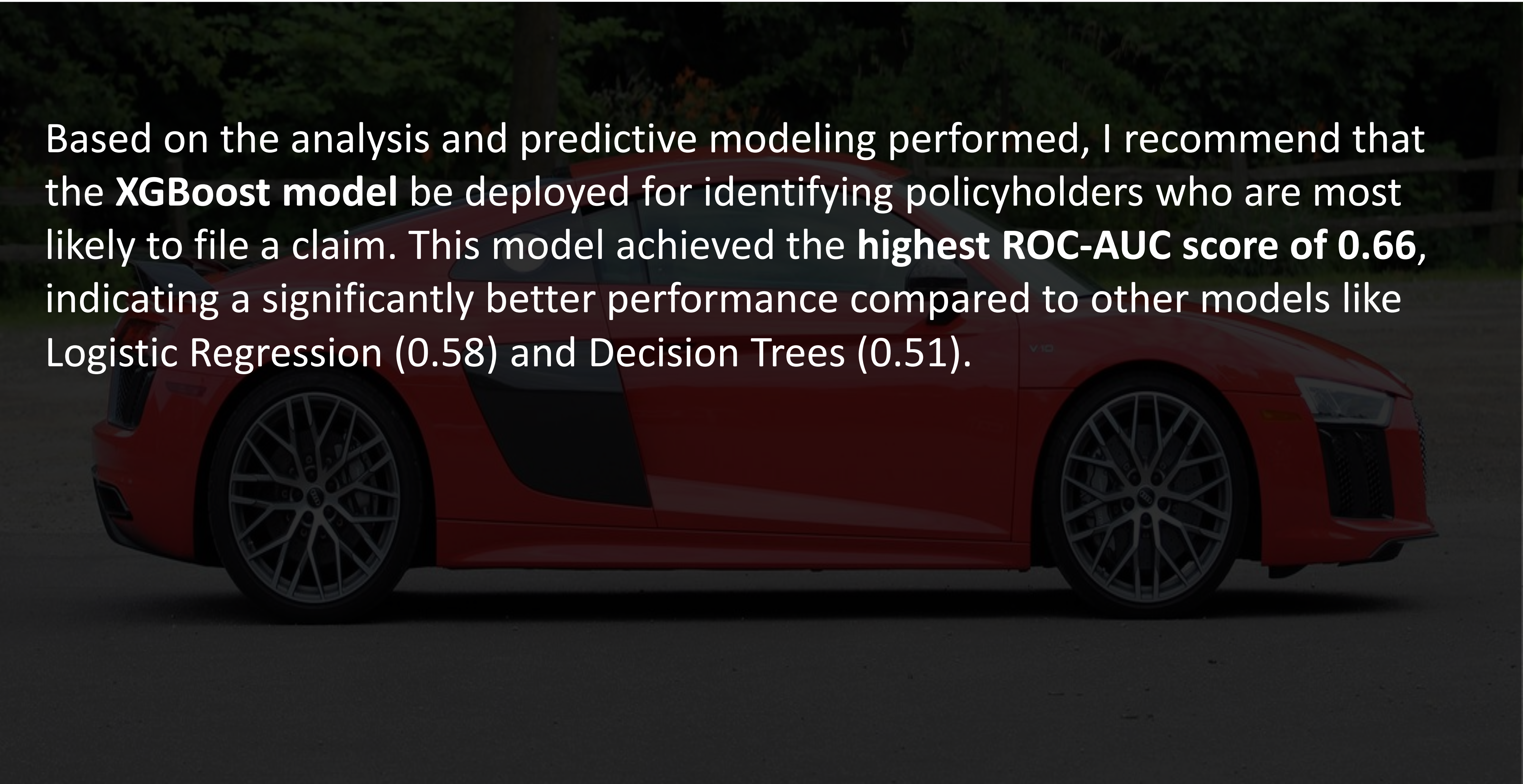
Insight: The ROC curve comparison plot shows that among all the models tested, XGBoost achieves the highest AUC score (0.66), indicating it is the best at distinguishing between policyholders who will and will not file a claim. Random Forest also performs well with an AUC of 0.65, showing strong predictive power. In contrast, Logistic Regression (AUC = 0.59) and especially the Decision Tree (AUC = 0.51) perform worse, with the Decision Tree barely outperforming random guessing

TOP 15 IMPORTANT FEATURES



Insight: From the best performing model we see that the top performing features are ‘age of car’, ‘height’, ‘policy tenture’ and ‘ncap rating’ have the strongest influence on whether a claim is predicted correctly

BUSINESS RECOMMENDATIONS



Based on the analysis and predictive modeling performed, I recommend that the **XGBoost model** be deployed for identifying policyholders who are most likely to file a claim. This model achieved the **highest ROC-AUC score of 0.66**, indicating a significantly better performance compared to other models like Logistic Regression (0.58) and Decision Trees (0.51).

RECOMMENDATIONS USING THE XGBOOST MODEL

Use the model to adjust premium

Customers with high predicted claim probability can be charged higher premiums

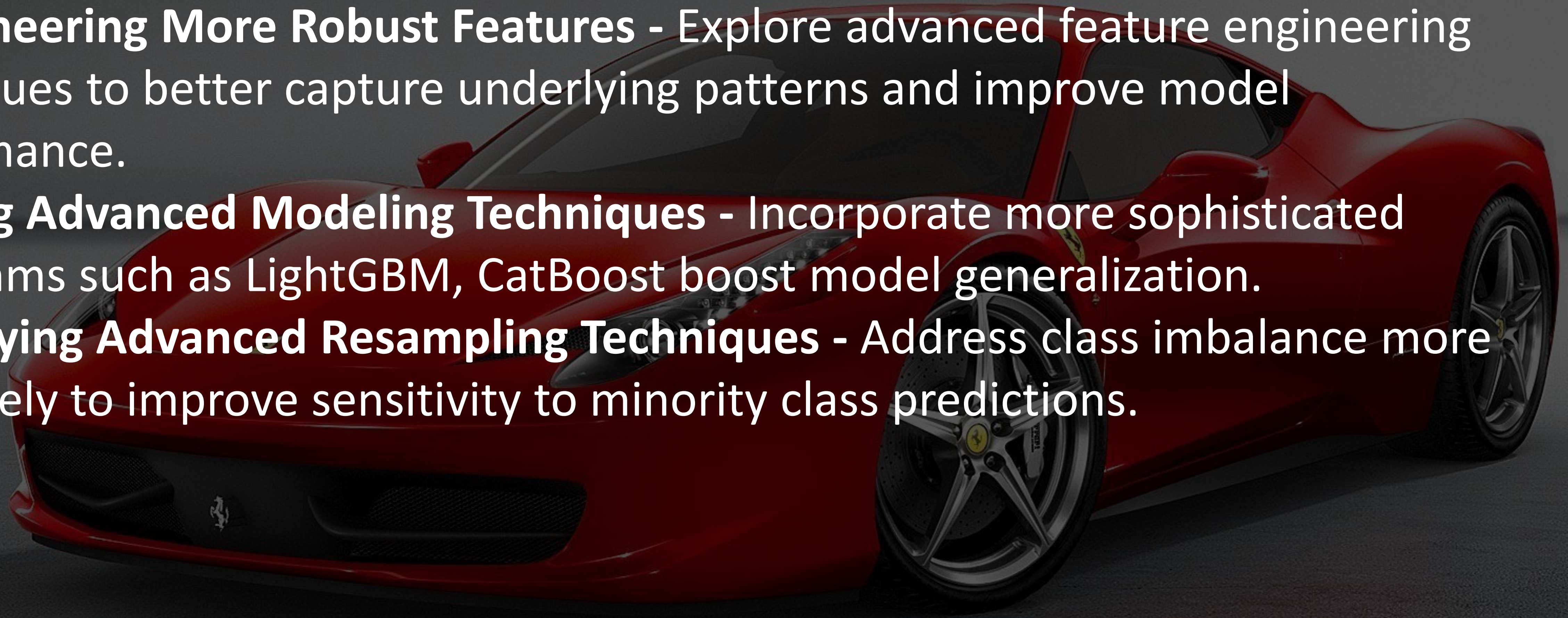
Improve fraud detection

Use the model's predictions to help **spot suspicious behavior** that could be used to enhance existing fraud detection systems

Offer proactive risk-reduction programs

Using the model to identify customers who are likely to file claims, then **offering them programs or guidance to reduce the risk before the claim even happens**

FUTURE WORKS



1. Engineering More Robust Features - Explore advanced feature engineering techniques to better capture underlying patterns and improve model performance.
2. Using Advanced Modeling Techniques - Incorporate more sophisticated algorithms such as LightGBM, CatBoost boost model generalization.
3. Applying Advanced Resampling Techniques - Address class imbalance more effectively to improve sensitivity to minority class predictions.

Thank You...!