CAR INSURANCE CLAIMS PREDICTION

By Group 2

Tishani Wijekoon(S16379), Chami Sewwandi(S16028), W.K.Hiruni Hasara(S16210), S.Luxan(s16329)

ABSTRACT

Accurately predicting car insurance claims is critical for risk assessment, pricing optimization, and financial stability in the insurance industry. This study focuses on developing a predictive model to assess the likelihood of a policyholder filing a claim within the next six months, using machine learning techniques such as XGBoost, Random Forest, Decision Tree, and SVM. The dataset, obtained from Kaggle, contains 58,592 observations across 44 variables, including policyholder demographics, policy characteristics, and vehicle attributes.

Exploratory Data Analysis (EDA) revealed that policy tenure, vehicle age, and policyholder age were the most significant predictors of claim occurrences. To improve model performance, feature selection and data balancing techniques (SMOTE and class weighting) were applied to address class imbalance, ensuring better claim detection. K-Means clustering was used for risk segmentation, categorizing policyholders into high, medium, and low-risk groups to support targeted pricing and policy adjustments. Among the models evaluated, XGBoost with SMOTE + Undersampling achieved the highest performance (92.34% accuracy, 98.68% recall), making it the most effective model for predicting claims and enhancing underwriting strategies. The results provide a data-driven framework for insurers to improve pricing accuracy, risk segmentation, and claims management, ultimately leading to better financial decision-making and customer satisfaction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Insurance companies aim to develop accurate and fair premium pricing models by assessing the risk levels of policyholders. One of the key challenges in the insurance industry is predicting the likelihood of a car insurance claim within a specific period. Effective risk assessment allows insurers to adjust policies, prevent fraud, and optimize financial stability while ensuring fair pricing for customers.

This study focuses on predicting car insurance claims within the next six months by analyzing policyholder demographics, policy characteristics, and vehicle attributes. Advanced machine learning techniques, including XGBoost, Random Forest, Decision Tree, SMV and clustering methods, are used to identify key risk factors and segment policyholders into risk groups. By distinguishing high-risk, medium-risk, and low-risk customers, insurers can implement dynamic pricing strategies, risk-adjusted policies, and proactive risk mitigation measures.

The findings from this study contribute to enhancing predictive modeling in the insurance industry, enabling insurers to refine risk assessment models, improve claim management efficiency, and develop personalized policy adjustments.

THE QUESTION WE ARE GOING TO ANSWER

"What key factors influence the likelihood of a car insurance claim being filed within the next six months?"

This study aims to identify the key factors influencing car insurance claims within the next six months and segment policyholders based on their risk levels. By analyzing policyholder demographics, policy characteristics, and vehicle attributes, we seek to understand:

- Which factors are most strongly associated with claim occurrences?
- How can policyholders be grouped into risk categories to optimize pricing and policy adjustments?

The findings from this study will support data-driven underwriting, improved risk management, and more accurate premium pricing strategies in the insurance industry.

DATA SET

The *"Car Insurance Claim Prediction"* dataset was obtained from Kaggle and consists of 58,592 observations across 44 variables (15 are numeric, while 29 are categorical).

Table 1- Dataset description

Categorical	
Variable	*Description*
policy_id	Policyholder's unique ID
area_cluster	Policyholder's area classification
make	Encoded manufacturer name
segment	Segment classification
model	Encoded name of the car
fuel_type	Type of fuel used
engine_type	Type of engine used
is_esc	Electronic Stability Control
is_adjustable_steering	Adjustability of steering wheel
is_tpms	Tyre Pressure Monitoring System
is_parking_sensors	Presence of parking sensors
is_parking_camera	Presence of parking cameras
rear_brakes_type	Type of rear brakes
transmission_type	Type of transmission
steering_type	Type of power steering
is_front_fog_lights	Presence of front fog lights
is_rear_window_wiper	Presence of rear window wiper
is_rear_window_washer	Presence of rear window washer
is_rear_window_defogger	Presence of rear window defogger
is_brake_assist	Availability of brake assist
is_power_door_locks	Presence of power door locks
is_central_locking	Availability of central locking
is_power_steering	Presence of power steering
is_driver_seat_height_adjustabl	Adjustability of driver seat height
is_day_night_rear_view_mirror	Presence of rearview mirror
is_ecw	Engine Check Warning Availability
is_speed_alert	Presence of Speed Alert System
ncap_rating	Safety Rating(out of 5)
is_claim	**Whether claim filed or not**

Numerical	
Variable	*Description*
policy_tenure	Time period of the policy
age_of_car	Normalized age of the car (years)
age_of_policyholder	Policyholder's Normalized age (yrs)
population_density	Policyholder's city density
displacement	Engine size (cc)
turning_radius	Space for completing a turn
length	Car length (mm)
width	Car width (mm)
height	Car height (mm)
gross_weight	Max allowable weight
max_torque	Maximum torque
max_power	Maximum power
airbags	Number of engine cylinders
cylinder	Number of airbags in the car
gear_box	Number of engine cylinders

IMPORTANT RESULTS OF THE DESCRIPTIVE ANALYSIS

Exploratory data analysis highlighted policy tenure, vehicle age, vehicle model, and geographical risk factors as the most significant predictors, while 38 out of 44 initial variables were removed due to redundancy, improving model efficiency.

We identified predicting insurance claims as class imbalance, with only 6.35% of policyholders filing claims *(Figure 1).* Policy tenure showed a strong association with claims *(Figure 2).* Geographical variations indicated slightly higher claim rates in low-density urban clusters, possibly due to differences in road infrastructure and accident frequency. Clustering analysis revealed overlapping risk segments (*Figure 3*, *Table 2*), highlighting the potential for improved feature selection and segmentation techniques to refine risk assessment strategies.

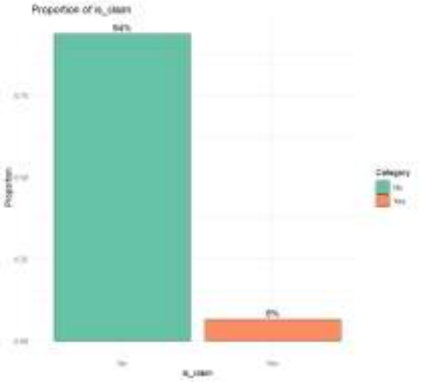


Figure 2-Proportion of response variable

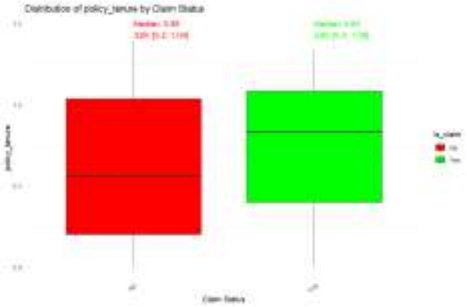


Figure 1-Distribution of policy tenure by claim status

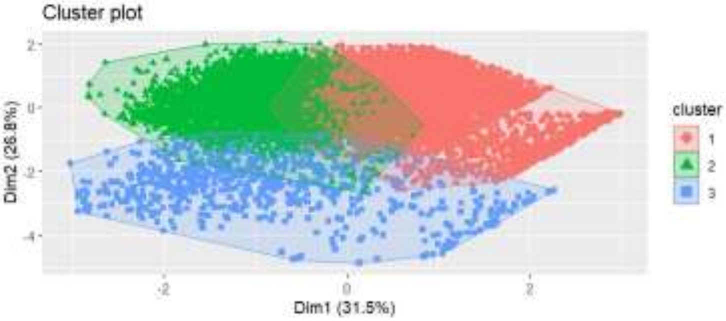
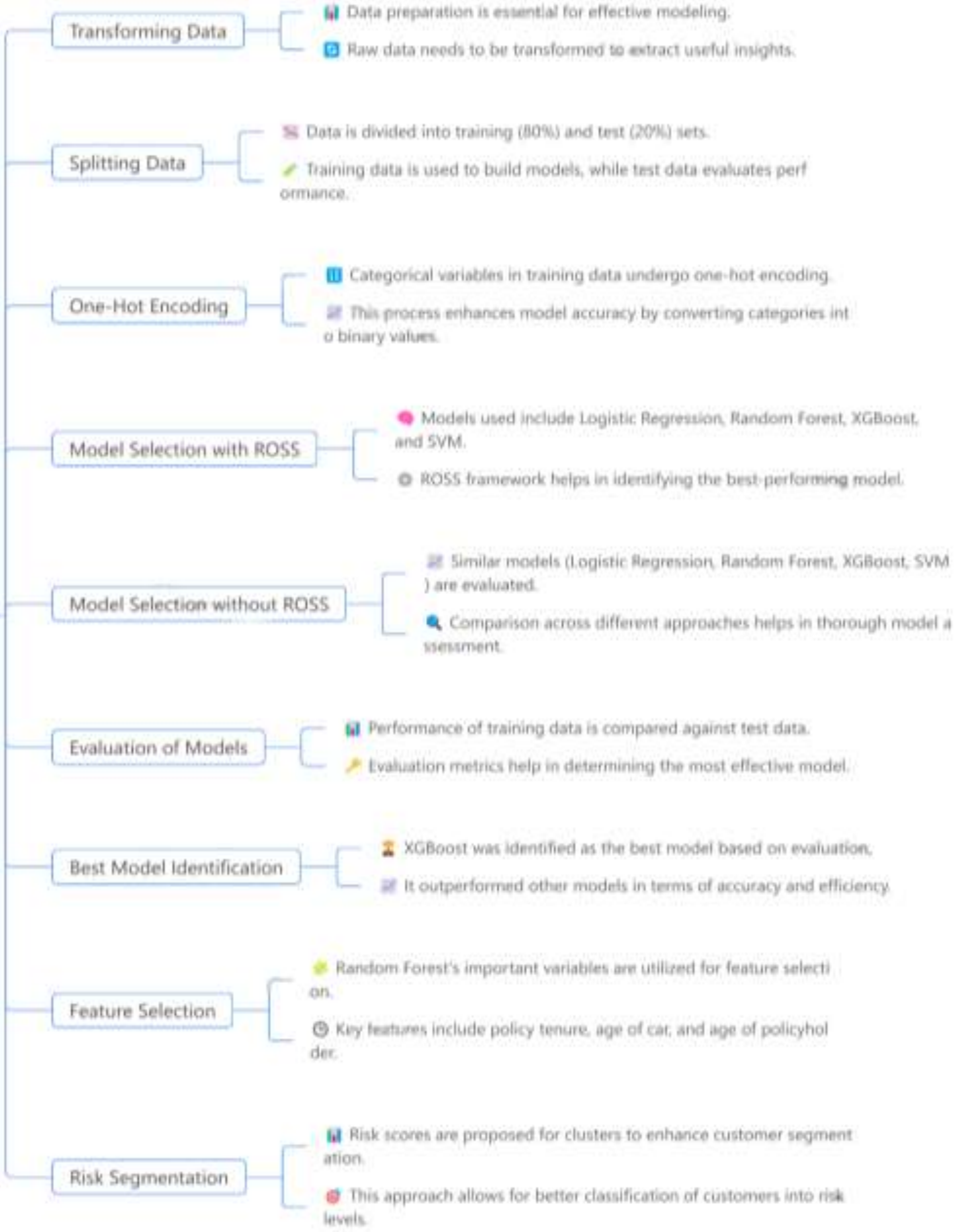


Figure 3-Cluster plot before the feature selection stage

Table 2- Summary Table for Kmean Clusters before feature selection

Risk_Level	Count	Avg_Age	Avg_Policy_Tenure	Avg_Car_Age	Claim_Rate
1	1468	0.259	0.621	0.408	0.062
2	2238	0.263	0.714	0.417	0.071
3	1294	0.267	0.411	0.209	0.062

IMPORTANT RESULTS OF THE ADVANCED ANALYSIS



Data preparation is essential for effective modeling.

Raw data needs to be transformed to extract useful insights.

Data is divided into training (80%) and test (20%) sets.

Training data is used to build models, while test data evaluates performance.

Categorical variables in training data undergo one-hot encoding.

This process enhances model accuracy by converting categories into binary values.

Models used include Logistic Regression, Random Forest, XGBoost, and SVM.

ROSS framework helps in identifying the best-performing model.

Similar models (Logistic Regression, Random Forest, XGBoost, SVM) are evaluated.

Comparison across different approaches helps in thorough model assessment.

Performance of training data is compared against test data.

Evaluation metrics help in determining the most effective model.

XGBoost was identified as the best model based on evaluation.

It outperformed other models in terms of accuracy and efficiency.

Random Forest's important variables are utilized for feature selection.

Key features include policy tenure, age of car, and age of policyholder.

Risk scores are proposed for clusters to enhance customer segmentation.

This approach allows for better classification of customers into risk levels.

Figure 4- Process of the advanced analysis

1. MODEL PERFORMANCE COMPARISON

To predict car insurance claims within the next six months, multiple machine learning models were evaluated using different balancing techniques. The models were assessed based on accuracy, precision, recall, and F1-score to determine the most effective approach for claim prediction. (Sahai, 2022)

1.1 Performance with SMOTE + Undersampling

Table 3- Model performance (with SMOTE + Undersampling)

Model	Dataset	Accuracy	Precision	Recall	F1
Logistic Regression	Training Set	58.49%	59.16%	54.85%	56.92%
	Test Set	55.19%	95.55%	54.57%	69.47%
Random Forest	Training Set	97.23%	98.79%	95.62%	97.18%
	Test Set	82.56%	94.08%	86.75%	90.27%
Decision Tree	Training Set	68.18%	73.75%	56.47%	63.96%
	Test Set	56.12%	95.06%	55.93%	70.42%
XGBoost	Training Set	93.52%	88.75%	99.68%	93.90%
	Test Set	92.34%	93.48%	98.68%	96.01%
SMV	Training Set	62.28%	64.13%	55.75%	59.64%
	Test Set	54.28%	96.09%	53.22%	68.50

1.2 Performance with Inbuilt Class Weights

Table 4- Model performance (with Inbuilt Class Weights)

Model	Dataset	Accuracy	Precision	Recall	F1
Random Forest	Training Set	94.21%	99.99%	93.82%	96.81%
	Test Set	88.05%	93.98%	93.16%	93.57%
Decision Tree (Pruned)	Training Set	62.89%	68.11%	48.47%	56.64%
	Test Set	49.57%	95.59%	48.06%	64.04%
XGBoost	Training Set	67.02%	98.07%	66.08%	78.96%
	Test Set	63.32%	95.1%	64.03%	76.53%

2. Model-Specific Analysis

Each model was analyzed based on its strengths, weaknesses, and suitability for claim prediction.

2.1 Logistic Regression

Logistic Regression serves as a baseline model to understand the relationship between predictor variables and claim probability. While it offers high interpretability, it struggles with complex non-linear patterns, leading to low accuracy (55.19%) on the test set. The model achieved a high precision of 95.55%, meaning that when it predicted a claim, it was mostly correct. However, its recall (54.57%) was low, indicating that a significant number of actual claims were misclassified as non-claims. This suggests that Logistic Regression fails to capture intricate relationships in the data, making it unsuitable for real-world insurance claim prediction, where missing a claim can be costly.

2.2 Random Forest

Random Forest performed significantly better than Logistic Regression, particularly in handling complex decision boundaries. When trained using SMOTE + Undersampling, it achieved a test accuracy of 82.56% and recall of 86.75%, ensuring better claim detection. However, the training accuracy was 97.23%, suggesting overfitting to the training data.

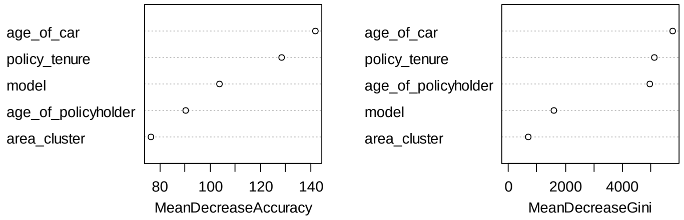


Figure 6-Variable importance plots of Random Forest with SMOTE + Undersampling

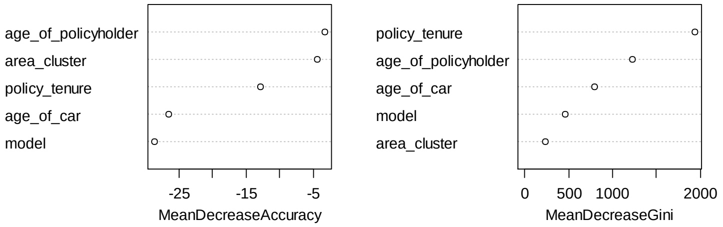


Figure 5- Variable importance plots of Random Forest with inbuilt class weights

To mitigate this, an alternative version using inbuilt class weights was tested, which achieved a better-balanced performance with a test accuracy of 88.05% while reducing overfitting. The feature importance analysis indicated that Age of Car, Policy Tenure, and Vehicle Model were the most influential variables. While Random Forest offers robust classification capabilities, it is computationally expensive and requires significant tuning to balance performance and generalization.

2.3 Decision Tree

The Decision Tree model (Decision Tree Pruning, n.d.) exhibited high precision (~95%) but poor recall (48-56%), meaning that while it was highly confident in its predictions, it missed a significant portion of actual claims. This led to poor generalization to unseen data, with a test accuracy of 56.12%. Additionally, overfitting was a major issue, as the training accuracy was significantly higher. To address this, a pruned version of the Decision Tree was tested, but it resulted in even lower accuracy due to an excessive reduction in model complexity. Given its tendency to overfit and its lower generalization ability, Decision Trees are not an ideal choice for claim prediction in a real-world insurance setting. (Trees and Classification, n.d.)

2.4 XGBoost

XGBoost with SMOTE + Undersampling was the best-performing model, achieving a test accuracy of 92.34% and recall of 98.68%, ensuring that nearly all claims were detected.

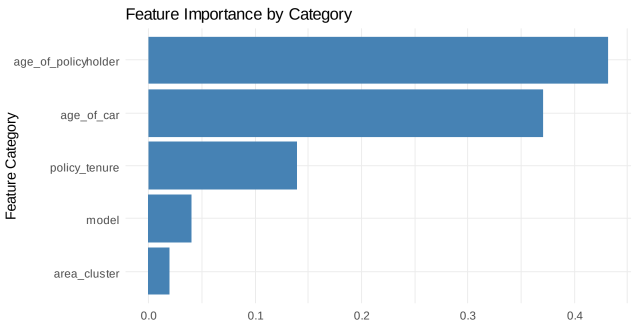


Figure 8-Variable importance plots of XGBoost with SMOTE + Undersampling

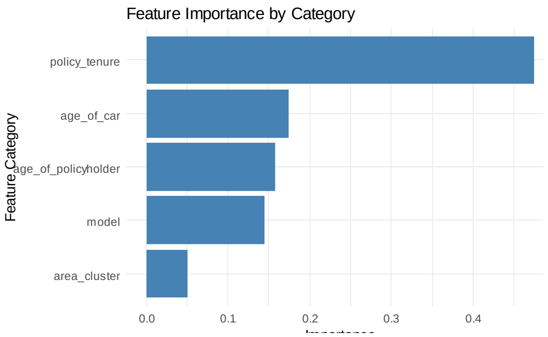


Figure 7- Variable importance plots of XGBoost with Inbuilt Class Weights

Unlike Decision Trees and Random Forest, XGBoost uses boosting techniques to iteratively improve model accuracy, making it more adaptable to complex patterns in the data. The variable importance analysis *(Figure 7,8)* showed that *Policy Tenure*, *Age of Car*, and *Age of Policyholder* were the most critical factors in predicting claims. While XGBoost slightly sacrificed precision (93.48%) compared to Random Forest, its exceptional recall ensures a minimal number of false negatives, reducing the risk of undetected claims. However, XGBoost requires careful hyperparameter tuning and is computationally intensive, making it more challenging to deploy at scale without optimization. (XGBoost, n.d.)

2.5 Support Vector Machine (SVM)

Support Vector Machine (SVM) (Support Vector Machine (SVM) Algorithm, n.d.) performed poorly compared to ensemble models. While it achieved a high precision of 96.09%, its recall was only 53.22%, indicating that it failed to identify a significant portion of actual claims. This suggests that SVM struggles with imbalanced datasets, as it is less effective in handling class disparities without additional resampling techniques. Unlike tree-based models, SVM is computationally expensive and scales poorly with large datasets, making it impractical for large-scale insurance claim prediction. Given these limitations, SVM is not a recommended approach for this problem.

3. RISK MANAGEMENT

The risk segmentation analysis provides a data-driven foundation for insurers to develop customized pricing models and targeted policy adjustments based on policyholder risk levels. By leveraging XGBoost, we identified the most influential variables for claim prediction: policy tenure, vehicle age, and policyholder age. Using these key features, we applied K-means clustering to segment policyholders into three distinct risk categories, summarized in *Table 5*

Table 5-Summary table for Kmean clusters after feature selection

Risk Level	Count	Avg policyholders age	Avg policy tenure	Avg car age	Claim rate
1	1661	0.2874	**0.997**	0.4159	**14.50%**
2	1172	0.2738	**0.2912**	0.1673	**7.84%**
3	2167	0.2393	**0.4818**	0.4238	**0**

Insurers can use risk segmentation to adjust policies and pricing based on risk levels, ensuring a balanced approach to risk management. High-risk customers should have stricter terms like higher deductibles and additional documentation to mitigate potential losses. Medium-risk customers require closer monitoring, possibly through telematics-based policies that track driving behavior. Low-risk customers can benefit from discounted premiums and loyalty rewards to encourage retention.

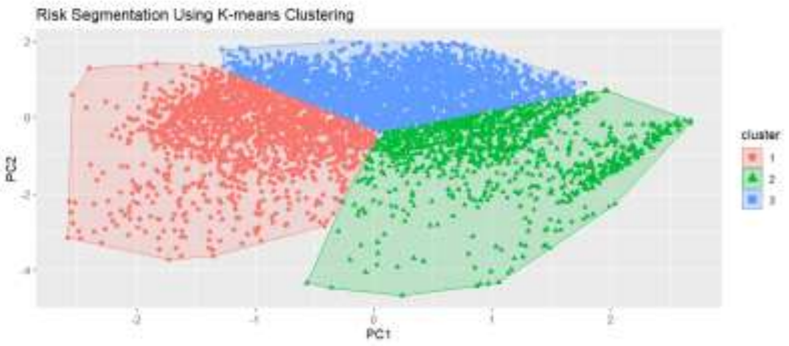


Figure 9-Risk segmentation using L-means clustering after feature selection

This segmentation helps insurers optimize pricing, reduce financial exposure, and improve customer retention. High-risk customers contribute proportionally to claim payouts, while medium-risk customers undergo adaptive monitoring. Low-risk customers receive incentives, fostering long-term engagement. Integrating risk segmentation into premium pricing models enhances profitability, fairness, and overall financial stability.

CHALLENGES FACED AND EFFECTIVE SOLUTIONS IMPLEMENTED

1. Severe Class Imbalance in Claim Data

The dataset exhibited a significant class imbalance, with only 6.35% of policyholders filing claims, while the majority (93.65%) had no claims. This imbalance led to biased model predictions, favoring the majority class and resulting in poor recall for actual claims. To address this, (5 Techniques to Handle Imbalanced Data For a Classification Problem, n.d.) Synthetic Minority Oversampling Technique (SMOTE) (SMOTE for Imbalanced Classification with Python, n.d.) combined with undersampling was applied to balance the dataset. This improved model recall without overfitting, ensuring that high-risk claims were detected more effectively. Additionally, cost-sensitive learning approaches (inbuilt class weights) were tested to provide an alternative way to handle imbalance.

2. Identifying the Most Relevant Risk Factors

Another key challenge was identifying the most relevant risk factors from the 42 available features. Many of these variables were found to be redundant with claim occurrence, making the models unnecessarily complex and harder to interpret. To optimize performance, a feature selection process was conducted using XGBoost's importance ranking. The analysis revealed that *Age of Car*, *Policy Tenure*, and *Policyholder Age* were the most influential predictors. Removing less significant variables enhanced both computational efficiency and interpretability, allowing insurers to focus on key risk indicators when designing policy adjustments.

3. Overfitting in Complex Models

Overfitting was another concern, particularly in complex models like Random Forest and XGBoost. While these models performed exceptionally well on training data, they showed reduced generalization to unseen test data. To counter this, hyperparameter tuning was performed by adjusting tree depth, learning rates, and regularization parameters. Additionally, comparing SMOTE-based oversampling with inbuilt class weighting allowed for the selection of a more generalized model, preventing overfitting while maintaining high accuracy.

4. Risk Segmentation with Meaningful Clusters

Risk segmentation was a critical challenge due to overlapping characteristics among policyholders, making it difficult to clearly distinguish between high, medium, and low-risk groups. Initial clustering attempts resulted in poor separation, which could lead to misclassification and ineffective policy adjustments. To enhance segmentation, K-Means clustering with optimized feature selection was applied, ensuring that the most relevant factors *vehicle age*, *policyholder age* and *Policy Tenure* were used for differentiation. Further refinements, such as adjusting cluster centroids and applying scaling techniques, improved separation, making the segmentation process more actionable for insurers.

DISCUSSION AND CONCLUSIONS

The analysis of various machine learning models for predicting car insurance claims within six months highlighted the effectiveness of **XGBoost with SMOTE + Undersampling**, which outperformed other models with **92.34% accuracy and 98.68% recall**. This model demonstrated the best balance between detecting actual claims and minimizing false positives, ensuring a robust and reliable risk assessment framework for insurers. Feature importance analysis identified **Age of Car, Policy Tenure, and Policyholder Age** as the strongest predictors, confirming that vehicle characteristics and policy duration play a critical role in claim likelihood. While Random Forest also performed well, it exhibited signs of overfitting, whereas Logistic Regression and SVM struggled with complex relationships in the data, making them less suitable for claim prediction.

These findings provide a data-driven foundation for risk segmentation and policy optimization in the insurance industry. By leveraging the clustering results, insurers can implement risk-adjusted pricing strategies, ensuring that high-risk customers are subjected to stricter policy terms, medium-risk customers are closely monitored, and low-risk customers receive incentives for long-term engagement. Future improvements should focus on incorporating external risk factors, refining fraud detection techniques, and leveraging real-time data analytics to enhance predictive accuracy. By integrating these insights into claims management and underwriting practices, insurers can achieve better financial stability, improved decision-making, and enhanced customer satisfaction.

APPENDIX

- R code
- **Precision** measures how many of the predicted claims were actually correct, ensuring fewer false positives.
- **Recall** indicates how many of the actual claims were correctly identified, reducing false negatives. and Recall.
- **Underwriting** is the process used by insurers to evaluate the risk of insuring a potential policyholder and determine the appropriate premium to charge.

BIBLIOGRAPHY

5 Techniques to Handle Imbalanced Data For a Classification Problem. (n.d.). Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/

Baran, S. a. (2022). Prediction of motor insurance claims occurrence as an imbalanced machine learning problem. *arXiv preprint arXiv:2204.06109*.

Decision Tree Pruning. (n.d.). Retrieved from KDnuggets: https://www.kdnuggets.com/2022/09/decision-tree-pruning-hows-whys.html

Sahai, R. a.-A.-H.-S. (2022). Insurance risk prediction using machine learning. In *The international conference on data science and emerging technologies* (pp. 419--433).

SMOTE for Imbalanced Classification with Python. (n.d.). Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

Support Vector Machine (SVM) Algorithm. (n.d.). Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

Trees and Classification. (n.d.). Retrieved from https://fderyckel.github.io/machinelearningwithr/trees-and-classification.html

XGBoost. (n.d.). Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/xgboost/