EXPLORATORY ANALYSIS OF CAR INSURANCE CLAIMS PREDICTION DATA

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

ABSTRACT

Car insurance claim prediction is essential for insurers to assess risk, refine pricing models, and improve decision-making. This study explores key factors influencing the likelihood of a policyholder filing a claim within the next six months by analyzing policyholder demographics, policy details, and vehicle specifications.

A comprehensive exploratory data analysis (EDA) was conducted on a dataset consisting of 58,592 observations and 44 variables. The preprocessing stage involved handling missing values, encoding categorical variables, removing redundant features, and selecting the most relevant attributes. Notably, car model was identified as a representative variable for all vehicle specifications, reducing redundancy and improving efficiency.

The findings from univariate, bivariate, and multivariate analyses highlight significant relationships between policy tenure, vehicle characteristics, and claim likelihood. Advanced statistical techniques, including Cramér's V, Factor Analysis of Mixed Data (FAMD), and correlation analysis, were applied to uncover key insights.

This study provides a data-driven foundation for enhancing risk assessment models and optimizing insurance pricing strategies, ultimately leading to fairer premiums and improved financial sustainability in the insurance sector.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Car insurance claim prediction plays a crucial role in helping insurers assess risk and refine their pricing strategies. Insurance companies consider various factors, such as how long a policy has been active, the type of car, and the policyholder's demographic information, to estimate the likelihood of claims. Understanding these factors allows insurers to adjust their pricing models, reduce financial risk, and offer fairer premiums to customers.

In this study, we'll take a closer look at a car insurance dataset to identify the main factors that influence whether a policyholder is likely to make a claim in the next six months. By exploring the data, we aim to uncover patterns that can inform more accurate, data-driven decisions, ultimately improving practices in the insurance industry.

DESCRIPTION OF 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 explore the key determinants of car insurance claims by analyzing various policyholder demographics, policy characteristics, and car specifications. By identifying patterns and relationships within these variables, we seek to understand:

- Which factors are most strongly associated with claim occurrences?
- Does the duration of a policy or demographic characteristics of the policyholder contribute to claim frequency?

Through this analysis, we aim to uncover key insights that can help insurance companies refine risk assessment models, improve premium pricing strategies, and enhance overall decision-making in the insurance industry.

DESCRIPTION OF THE 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).

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

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)

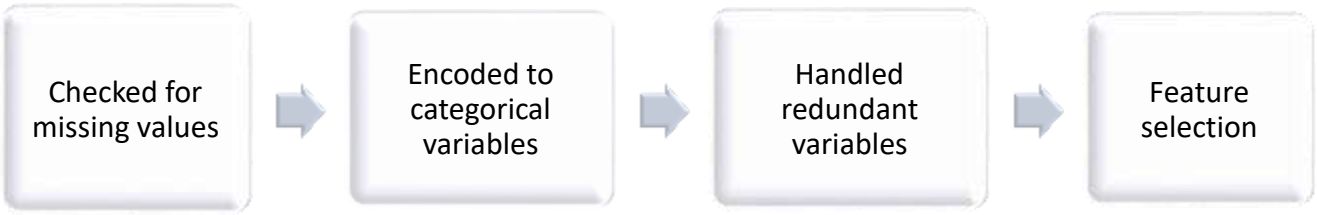
fuel_type	Type of fuel used
engine_type	Type of engine used
is_esc	Electronic Stability Control
is_adjustable_steering	Adjustability of steering wheel
is_tpms	Tyre Pressure Monitoring System
is_parking_sensors	Presence of parking sensors
is_parking_camera	Presence of parking cameras
rear_brakes_type	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**

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 airbags in the car
cylinder	Number of engine cylinders
gear_box	Number of gears in the car

Table 1-Description of the data set

DATA PREPROCESSING

Before conducting exploratory data analysis, we verified that the dataset was pre-split into training and test sets. The training set consists of 58,592 observations and is used for data cleaning, feature selection, and analysis.



- Checking for missing values, duplicates and outliers
  - Missing values- none found
  - Duplicates- none found
  - Outliers- didn't remove since they might explain some inherent variability attributed with claiming

- Encode to categorical variables
  - "area_cluster" variable was recategorized based on population density - "Low", "Medium", "High"

- Handling redundant variables
  - Removed "population_density" variable (*area_cluster* and *population_density* provide the same info)
  - Removed "length", "width", "height", and "gross_weight" (provide the same info that is given by "segment") (All Car Segments Types In India Explained With Examples, n.d.)

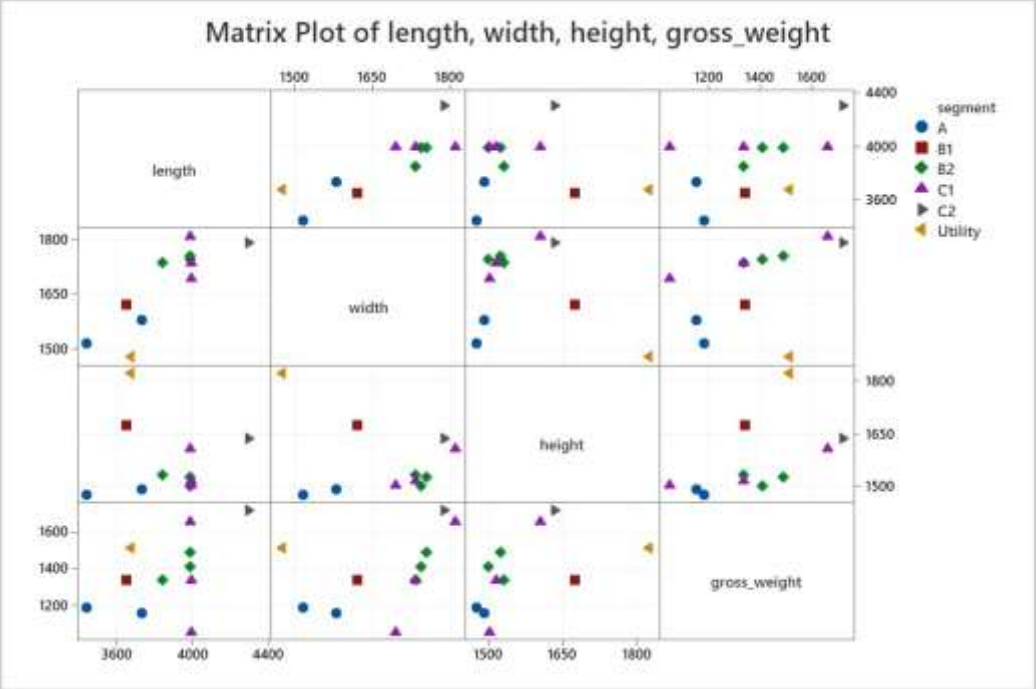


Figure 1-Segment vs length, width, height, gross_weight
(We observed "length", "width", "height", and "gross_weight" provide the same info that is given by "segment")

Evaluating the redundancy of the "length," "width," "height," and "gross weight";

We applied Cramér's V to assess the association between "segment" and "length," "width," "height," and "gross_weight". Since Cramér's V values are very close to 1, the results (Figure 2) indicate a perfect association, meaning these specifications are redundant as they provide the same information as the "segment" variable.

```
> # Print results
> print(c(cramer_length, cramer_width, cramer_height, cramer_weight))
[1] 0.9525747 0.9586968 0.9934882 0.9754621
> |
```

Figure 2-Evaluating the redundancy of the terms "length," "width," "height," and "gross weight"

— We observed each car model has a unique set of specifications, including individual attributes (such as fuel type, engine power, safety features, and transmission type) would be redundant. Instead, the model variable serves as a representative feature for all car specifications.

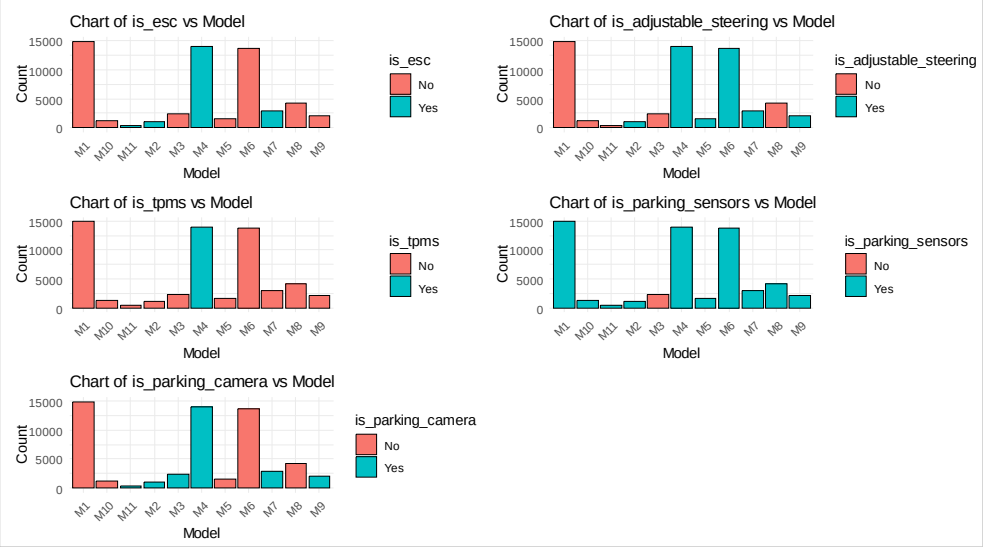


Figure 3-A few bar charts that describe each car model have a unique set of specifications

Evaluating the redundancy of car specification variables;

We used Cramér's V to measure the association between "model" and car specifications. *(What does Cramer's V actually measure and tell us?, n.d.)*

With a value of 1 for each pair, the results *(Figure 4)* indicate a perfect association, meaning the car specifications are entirely dependent on "model" and are redundant. In other words, the "model" variable contains complete information about the car specifications, making the two variables strongly related.

```
> # View results
> print(cramer_results)
                        make                        segment
                           1                              1
                       model                      fuel_type
                           1                              1
                 engine_type                        airbags
                           1                              1
            rear_brakes_type                       cylinder
                           1                              1
              transmission_type                       gear_box
                           1                              1
                steering_type                     ncap_rating
                           1                              1
                   max_torque                      max_power
                           1                              1
                      is_esc          is_adjustable_steering
                           1                              1
                     is_tpms                is_parking_sensors
                           1                              1
           is_parking_camera              is_front_fog_lights
                           1                              1
          is_rear_window_wiper            is_rear_window_washer
                           1                              1
       is_rear_window_defogger                 is_brake_assist
                           1                              1
          is_power_door_locks                is_central_locking
                           1                              1
        is_power_steering  is_driver_seat_height_adjustable
                           1                              1
is_day_night_rear_view_mirror                         is_ecw
                           1                              1
                is_speed_alert                   displacement
                           1                              1
```

Figure 4-Output of Cramer'V test

- Feature selection
  - Removed features that are not useful (policy ID, which has no predictive value).
  - Identify important variables: we selected demographic factors (age_of_policyholder, area_cluster), policy tenure, age of the car, and model of the car as key variables.

THE MAIN RESULTS OF THE DESCRIPTIVE ANALYSIS

FROM UNIVARIATE ANALYSIS;

This step is crucial to identify potential issues such as skewed distributions or outliers, which may affect the model performance.

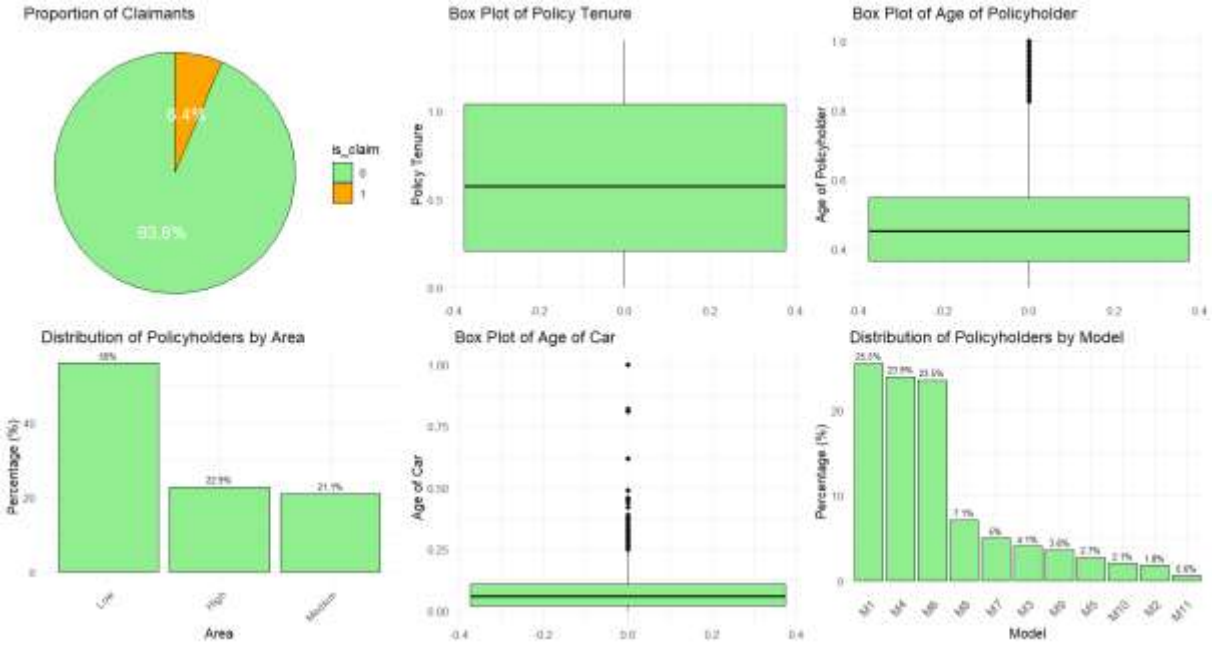


Figure 5-Univariate Analysis

- Response - The **claim status** distribution shows that **94% of policyholders did not file a claim,** while only **6% did.** This imbalance suggests that accidents or claim-worthy incidents are relatively rare.
- Policy Tenure - The boxplot of policy tenure shows a **median around 0.6**, suggesting that most policy tenures are concentrated around this value. The interquartile range, represented by the box's length, indicates the spread of the middle 50% of tenures.
- Age of Policyholder - The normalized age of policyholders shows **a right-skewed distribution** with most values on the lower side and some **high-end outliers**. This suggests a younger customer base, likely due to affordability, policy benefits, or employer-provided insurance.
- Area of Policyholder - There is a **dominance of the "low" category** at 56%, suggesting a potential concentration of policies in less populated regions. Probable reasons are factors like varying living costs, population density, or targeted marketing efforts.

- Age of Car - With a distribution **skewed toward higher values**, it is safe to say some cars seem to be relatively old. This can be due to insurance policies catering to long-term car owners, while newer cars may be underrepresented.

- Model – There is a clear preference for **'M1' ,'M4', and 'M6'** accounting for **about 75%** possibly due to their features, affordability, or brand image. The lower preference may be due to factors like higher price points, limited availability, or less appealing features.

FROM BIVARIATE ANALYSIS;

Demographic Factors vs Response

- Age of Policyholder vs Response

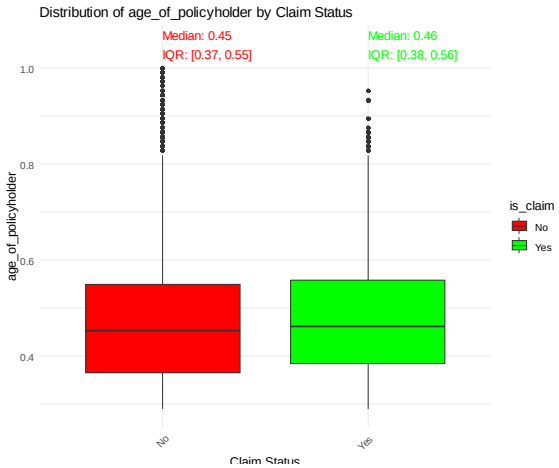


Figure 6-Boxplot of age_of_policyholder by claim Status

The boxplot compares the age distribution of policyholders based on their claim status. The median age for both claim (Yes) and non-claim (No) groups is nearly identical, with 45 years for non-claim cases and 46 years for claim cases, suggesting that policyholder age alone does not significantly influence claim likelihood. The similarity in distributions may result from balanced risk profiles across age groups. Despite common beliefs, younger and older drivers show similar claim patterns, suggesting that factors other than age, such as driving conditions or vehicle type, may have a stronger influence on claim frequency.

- Area of Policyholder vs Response

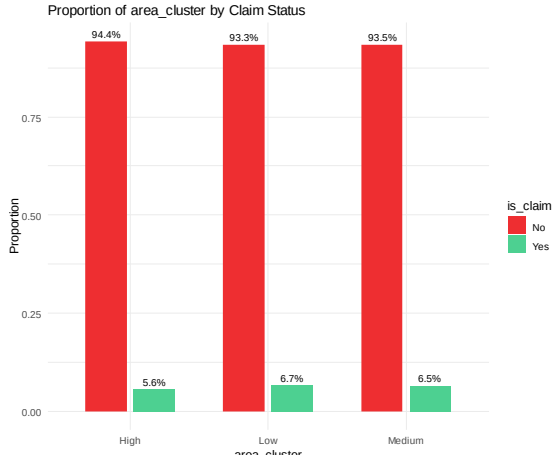


Figure 7--Bar chart of area_cluster by claim Status

The claim percentages are 5.6% for high-density areas, 6.5% for medium-density areas, and 6.7% for low-density areas. This shows that claim rates slightly increase as the population density decreases. One reason for this could be that areas with higher population density might have more competition among insurers, leading to better risk management and lower claim rates. On the other hand, low-density areas may experience less competition, which could result in slightly higher claim rates. Additionally, people in lower-density areas may have limited access to healthcare or fewer insurance options, which could influence the number and frequency of claims.

Car Specifications vs Response

- Age of Car vs Response

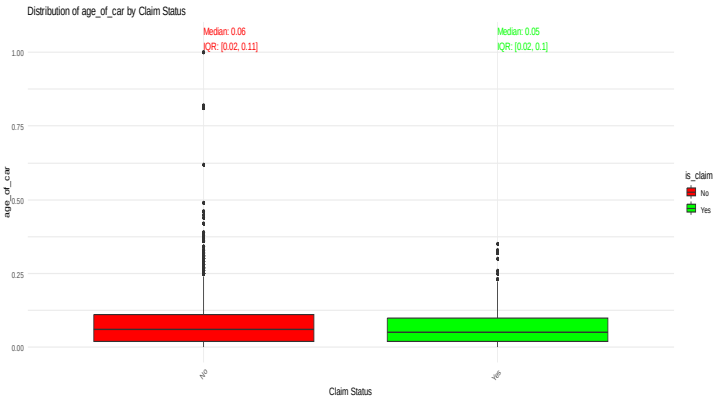


Figure 8-Boxplot of age_of_car by claim Status

The boxplot illustrates the distribution of car age for vehicles that filed an insurance claim (Yes) and those that did not (No). Both distributions appear nearly identical, with median car ages of **0.06 years for non-claim cases** and **0.05 years for claim cases**, along with similar interquartile ranges. This suggests that car age alone does not significantly influence claim status. The similarity in distributions could be due to several factors, such as **consistent maintenance across different car ages, balanced insurance coverage policies.**

Additionally, both new and old cars may have reasons for claims, newer cars might file claims due to accidental damage, while older cars might claim repairs.

- Model of car vs Responses

In the analysis, we observe that the response variable "is_claim" has different percentages across the 11 models, reflecting varying claim frequencies for each model. To better understand these patterns, we categorized the models based on the percentage of claims they represent, grouping them as follows:

- Models with less than 4.5% claims (M11)
- Models with claims between 4.5%-6.0% (M1, M3, M8, M10)
- Models with claims between 6.0%-7.5% (M4, M5, M6, M7, M9, M2)

The comparison of models based on their claim frequencies reveals varying levels of risk, with Model 11 (M11) showing the lowest claim percentage at 4.1%, indicating a lower risk of claims, while Models 2 (M2) and others in the 6.0%-7.5% range exhibit a higher risk. This suggests that certain models have specific car characteristics, that make them more prone to claims. The importance of comparing *'Model vs. is_claim'* is that it uncovers the relationships between car characteristics and claim frequencies, highlighting the specific risk factors that drive insurance claims.

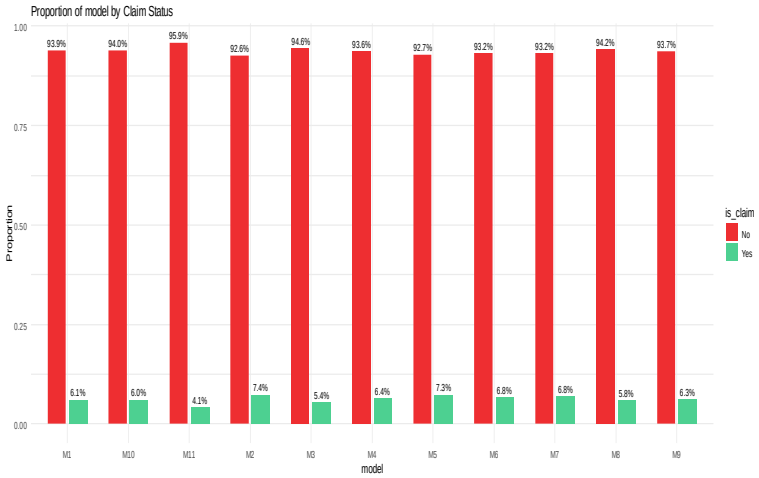
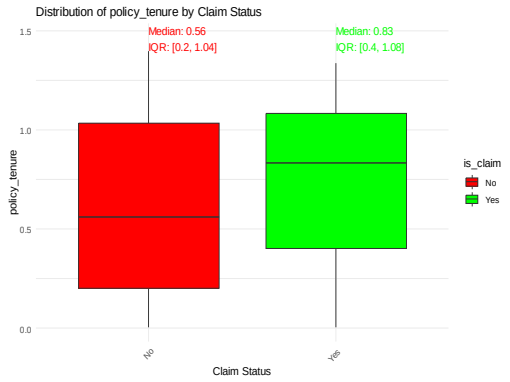


Figure 9-Bar chart of model by claim Status

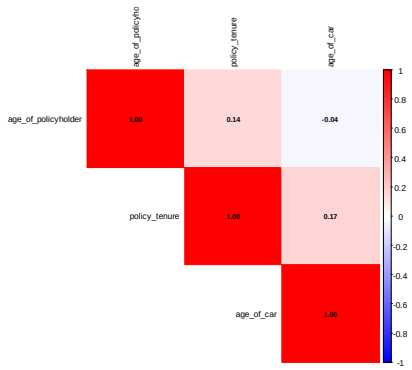
Policy Tenure vs Response



The median tenure is higher for claims, suggesting longer-held policies are more likely to file claims. The interquartile range is also higher for claims, indicating fewer short-tenure policies making claims. This pattern may reflect policyholder commitment, where longer tenure increases the likelihood of experiencing and reporting incidents. Conversely, non-claims show greater variability, possibly due to policy cancellations, or lower-risk customers maintaining policies without needing claims.

Figure 10-Boxplot of policy_tenure by claim Status

Correlation Heat map of Numerical Variables



The *age_of_policyholder* and *policy_tenure* have a weak positive correlation of **0.14**, indicating that older policyholders tend to have slightly longer policy tenures. The *policy_tenure* and *age_of_car* show a weak positive correlation of **0.17**, suggesting that policy tenure slightly increases as the car's age increases. However, the *age_of_policyholder* and *age_of_car* have a weak negative correlation of **-0.04**, meaning there is little to no relationship between them.

Figure 11-Heatmap of numerical variables

FROM MULTIVARIATE ANALYSIS;

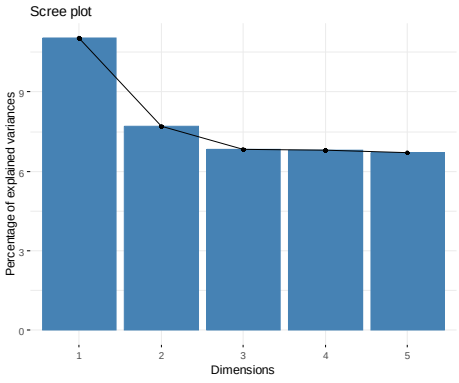


Figure 12- Scree plot of FACD

Since the dataset is a mixed dataset with continuous and categorical variables, FAMD (factor analysis of mixed data) was used. FAMD can be seen as combining PCA for continuous variables and multiple correspondence analysis (MCA) for categorical variables.

The scree plot displays the explained variance for each principal component obtained from the FAMD analysis. The x-axis represents the principal components, while the y-axis indicates the percentage of total variance explained by each component. The first principal component explains 11.040% of the variance, making it the most significant dimension. The

second component accounts for 7.719% of the variance, followed by the third component at 6.830%. The "elbow" point, where the curve starts to flatten, is observed at the 3rd component. This indicates that the first three components capture only 25.590% of the variability and that is not enough for further analysis based on these components.

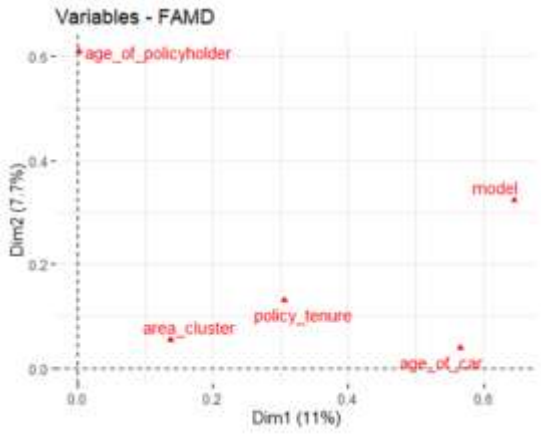


Figure 13– Variable factor map of FAMD

The **Variable Factor Map** from the FAMD indicates how the variables contribute to the first two dimensions, with **Dim 1 (11.040%)** explaining more variance than **Dim 2 (7.719 %)**. Variables such as **model** and **age_of_car** are strongly associated with Dim 1, suggesting they share similar patterns or contribute significantly to this dimension. In contrast, **age_of_policyholder** aligns closely with Dim 2 and weakely associated with Dim 1 suggesting that the variable is different from the other variables. **Policy_tenure** appears moderately correlated with both dimensions. **area_cluster** is weakly associated with the two dimensions.

SUGGESTIONS FOR A QUALITY ADVANCED ANALYSIS

To enhance the depth of this analysis, advanced statistical and machine learning techniques were considered. Logistic Regression with interaction effects, Random Forest, and XGBoost classification models were proposed to evaluate predictive power and feature importance. Additionally, SHAP value analysis was recommended to improve model interpretability.

For a more detailed understanding of risk, k-Means Clustering was used to segment policyholders based on claim likelihood. Furthermore, Bayesian Inference was suggested for claim probability estimation, offering a robust approach to modeling uncertainty and refining predictions.

By integrating these techniques, this study provides a comprehensive risk assessment framework that enhances insurance pricing, policy design, and fraud detection.

APPENDIX INCLUDING R CODE

EDA_R code-project1

BIBLIOGRAPHY

All Car Segments Types In India Explained With Examples. (n.d.). Retrieved from www.v3cars.com: https://www.bing.com/search?q=segment+of+the+car%28a%21%2Ca2%2Cb1%2Cb2%2Cc1%2Cc2%29&form=ANNTH1&refig=a7b4b52754a6474d8a55fb045ec14cc8&pc=W099&pq=seg&pqlth=3&assgl=37&sgcn=segment+of+the+car%28a%21%2Ca2%2Cb1%2Cb2%2Cc1%2Cc2%29&qs=HS&smvpcn=0&swbcn=10&sctcn

What does Cramer's V actually measure and tell us? (n.d.). Retrieved from www.reddit.com: https://www.reddit.com/r/explainlikeimfive/comments/a0igip/eli5_what_does_cramers_v_actually_measure_and/?rdt=52237

Baran, S. (n.d.). Prediction of motor insurance claims occurrence as an imbalanced machine learning problem.

Anwar, A. (2023). Predicting Likelihood That A Policyholder Will File A Claim In The Next Six Months. *Medium.* Retrieved from:

https://medium.com/@ammasanaswar/predicting-likelihood-that-a-policyholder-will-file-a-claim-in-the-next-six-months-c5a322446505

Reddy, Y. U. (2023). *Car Insurance Claim Prediction Model*. *Medium*. Retrieved from:

https://medium.com/@yudayreddy1/car-insurance-claim-prediction-model-bff7f06a9a77

GeeksforGeeks. (2023). *What is Exploratory Data Analysis?* Retrieved from:

https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/

Maklin, C. (2023). Isolation Forest. *Medium.* Retrieved from:

https://medium.com/@corymaklin/isolation-forest-799fceacdda4

Aggarwal, C. C. (2017). Outlier Analysis (2nd ed.). *Springer.*