CAR INSURANCE CLAIM PREDICTION

Abstract:

The project aims to create a predictive model that predicts a policyholder filing an insurance claim within six months. This model could revolutionize insurance companies resource management, risk assessment, and claims processing workflows. The model uses a comprehensive dataset of policyholder attributes, allowing for more accurate premium rate adjustments, improved claims processing efficiency, and effective risk mitigation strategies. The model will use Logistic Regression, Linear Discriminant Analysis, and Random Forest Classification for modeling binary outcomes and handling complex interactions. The insights will provide data-driven strategies, increased operational efficiency, and customer satisfaction.

1. INTRODUCTION:

The goal of this project is to create a simple tool that can be used to determine a customer's likelihood of filing a claim on their auto insurance during the next six months. We're looking into the experiences and details of nearly 59,000 customers. This covers fundamental information such as age, place of residence, length of time they've had insurance, and kind of vehicle they drive with a particular emphasis on the safety features of these vehicles. Our primary objective is to give insurance companies a more comprehensive view of possible claims. They will be better able to anticipate who would make a claim and prepare as a result. This project has two advantages. First, by seeing patterns and trends in claims, it assists insurance companies in better risk management. Second, as insurance providers will have a greater grasp of each client's unique circumstances, it may result in more individualized insurance policies for clients. To create this model, we're utilizing a number of straightforward but efficient techniques. To make sure our model is trustworthy and produces correct forecasts, we will extensively test it. By doing this, we want to produce a useful tool that insurance firms may utilize daily.

2. DATASET:

This dataset was sourced from Kaggle which contains information on 58,592 policyholders with 44 attributes, including policy, tenure, car and policyholder age, area cluster, car specifications, safety features, and more. The target variable is "is_claim" indicating whether a policyholder will file a claim in the next 6months or not.

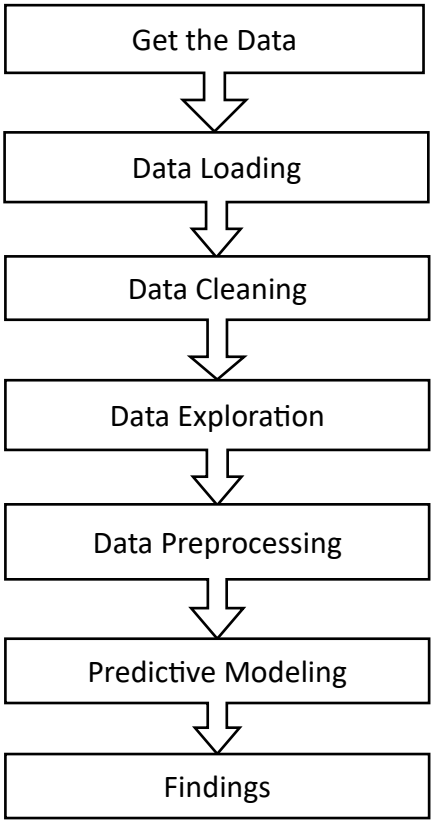
Information on the individual attributes can be found in the table below:

Variable	Description
policy_id	Unique identifier of the policyholder
policy_tenure	Time period of the policy
age_of_car	Normalized age of the car in years
age_of_policyholder	Normalized age of policyholder in years
area_cluster	Area cluster of the policyholder
population density	Population density of the city (Policyholder City)
make	Encoded Manufacturer/company of the car
segment	Segment of the car (A/ B1/ B2/ C1/ C2)
model	Encoded name of the car
fuel_type	Type of fuel used by the car
max_torque	Maximum Torque generated by the car (Nm@rpm)
max_power	Maximum Power generated by the car (bhp@rpm)
engine_type	Type of engine used in the car
airbags	Number of airbags installed in the car
is_esc	Boolean flag indicating whether Electronic Stability Control (ESC) is present in the car or not.
is_adjustable_steering	Boolean flag indicating whether the steering wheel of the car is adjustable or not.
is_tpms	Boolean flag indicating whether Tyre Pressure Monitoring System (TPMS) is present in the car or not.
is_parking_sensors	Boolean flag indicating whether parking sensors are present in the car or not.
is_parking_camera	Boolean flag indicating whether the parking camera is present in the car or not.
rear_brakes_type	Type of brakes used in the rear of the car
displacement	Engine displacement of the car (cc)
cylinder	Number of cylinders present in the engine of the car
transmission_type	Transmission type of the car
gear_box	Number of gears in the car
steering_type	Type of the power steering present in the car
turning_radius	The space a vehicle needs to make a certain turn (Meters)
length	Length of the car (Millimetre)
width	Width of the car (Millimetre)
height	Height of the car (Millimetre)
gross_weight	The maximum allowable weight of the fully-loaded car, including passengers, cargo and equipment (Kg)
is_front_fog_lights	Boolean flag indicating whether front fog lights are available in the car or not.
is_rear_window_wiper	Boolean flag indicating whether the rear window wiper is available in the car or not.
is_rear_window_washer	Boolean flag indicating whether the rear window washer is available in the car or not.
is_rear_window_defogger	Boolean flag indicating whether rear window defogger is available in the car or not.
is_brake_assist	Boolean flag indicating whether the brake assistance feature is available in the car or not.
is_power_door_lock	Boolean flag indicating whether a power door lock is available in the car or not.
is_central_locking	Boolean flag indicating whether the central locking feature is available in the car or not.
is_power_steering	Boolean flag indicating whether power steering is available in the car or not.
is_driver_seat_height_adjustable	Boolean flag indicating whether the height of the driver seat is adjustable or not.
is_day_night_rear_view_mirror	Boolean flag indicating whether day & night rearview mirror is present in the car or not.
is_ecw	Boolean flag indicating whether Engine Check Warning (ECW) is available in the car or not.
is_speed_alert	Boolean flag indicating whether the speed alert system is available in the car or not.
ncap_rating	Safety rating given by NCAP (out of 5)
is_claim	Outcome: Boolean flag indicating whether the policyholder file a claim in the next 6 months or not.

Figure 1: Description of the Attributes.

3. DATAFLOW AND ARCHITECTURE:

The data flow diagram below provides a visual representation of the data utilized in this project.



Data loading involves importing data into an analysis environment, followed by data cleaning, data exploration, data preprocessing, predictive modeling, and findings. Data cleaning removes inaccurate records, while data exploration identifies patterns and outliers. Preprocessing prepares data for modeling, while predictive modeling uses statistical or machine learning algorithms.

4. DATA CLEANING:

The data cleaning process involved meticulous checking for null values and simplifying the dataset by removing statistically significant variables, thereby enhancing the robustness and efficiency of subsequent analysis.

```
# A tibble: 6 × 44
  policy_id policy_tenure age_of_car age_of_policyholder area_cluster population_density  make segment model
  <chr>            <dbl>      <dbl>                <dbl> <chr>                      <dbl> <dbl> <chr>   <chr>
1 ID00001          0.516       0.05                0.644 C1                          4990     1 A       M1
2 ID00002          0.673       0.02                0.375 C2                         27003     1 A       M1
3 ID00003          0.841       0.02                0.385 C3                          4076     1 A       M1
4 ID00004          0.900       0.11                0.433 C4                         21622     1 C1      M2
5 ID00005          0.596       0.11                0.635 C5                         34738     2 A       M3
6 ID00006          1.02        0.07                0.519 C6                         13051     3 C2      M4
# i 35 more variables: fuel_type <chr>, max_torque <chr>, max_power <chr>, engine_type <chr>, airbags <dbl>,
#   is_esc <chr>, is_adjustable_steering <chr>, is_tpms <chr>, is_parking_sensors <chr>,
#   is_parking_camera <chr>, rear_brakes_type <chr>, displacement <dbl>, cylinder <dbl>,
#   transmission_type <chr>, gear_box <dbl>, steering_type <chr>, turning_radius <dbl>, length <dbl>,
#   width <dbl>, height <dbl>, gross_weight <dbl>, is_front_fog_lights <chr>, is_rear_window_wiper <chr>,
#   is_rear_window_washer <chr>, is_rear_window_defogger <chr>, is_brake_assist <chr>,
#   is_power_door_locks <chr>, is_central_locking <chr>, is_power_steering <chr>, …
```

Figure 2: Data set After Cleaning.

5. **DATA EXPLORATION:**

1) What is the distribution of the target variable?

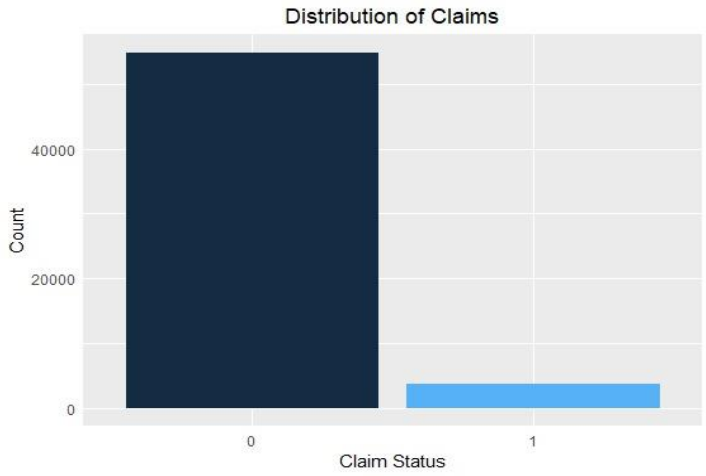


Figure 3: Distribution of Claims.

The visualization shows a bar chart displaying the distribution of a binary target variable in a dataset, with two categories marked by "0" and "1". The majority of data falls into category "0," indicating non-events or default status, while category "1" occurs less frequently and may reflect an event of interest, such as filed claims.

2) Does the preference for fuel type change with the age of the car and the age of the policyholder?

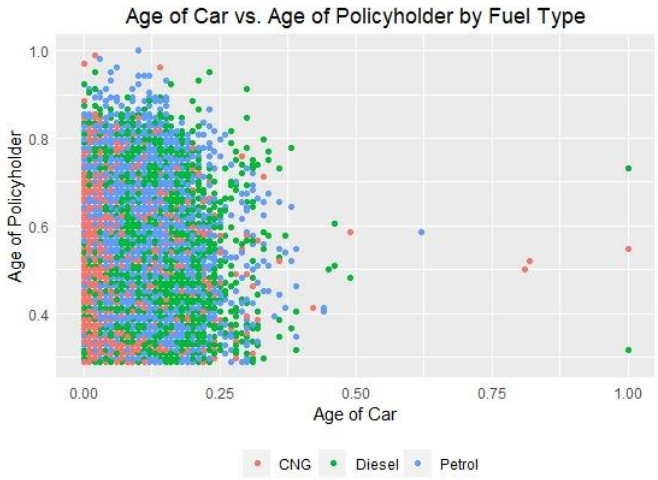


Figure 4: Age of Car vs Age of Policyholder by Fuel Type.

The scatterplot shows a comparison of car and policyholder age, categorized by fuel type: CNG, Diesel, and Petrol. The plot shows a diverse preference for fuel types across different ages, with no clear trend linking car or policyholder age to a specific fuel type. The distribution appears even, suggesting that fuel preference may not be significantly influenced by car or policyholder age within the dataset.

3) Is there a trend indicating that policyholders of certain ages prefer cars with a specific transmission type?

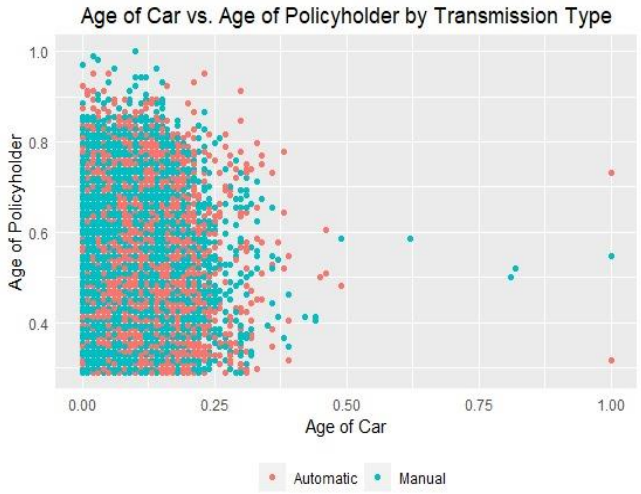


Figure 5: Age of Car vs. Age of Policyholder by Transmission Type.

The scatterplot shows the relationship between car and policyholder age, categorized by transmission type. There is no significant trend in preference for transmission type, but manual cars are more prevalent among younger vehicles. Automatic transmission cars are evenly distributed across all ages, while manual cars are fewer in older cars.

4) Which transmission type is most common among different age groups of policyholders?

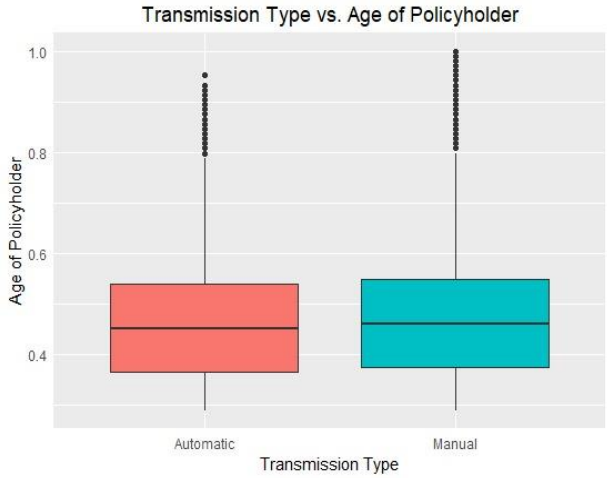


Figure 6: Transmission Type vs. Age of Policyholder.

The boxplot compares policyholders' ages and preferred transmission types, Automatic in red and Manual in cyan. Both categories have a similar median age and interquartile range, suggesting no significant age difference. Outliers indicate younger or older policyholders. The plot does not show a clear preference for one transmission type among different age groups.

5) What is the average age of policyholders within each area cluster?

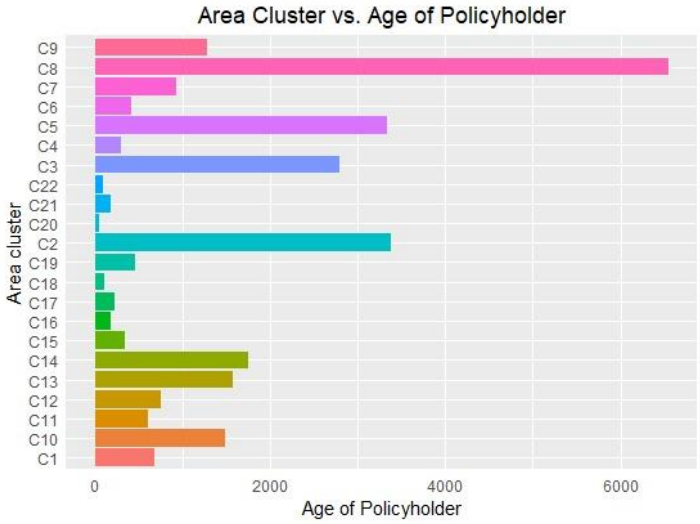


Figure 7: Area Cluster vs. Age of Policyholder.

The bar chart shows the average age of policyholders across different geographical clusters, with longer bars indicating higher averages in clusters like C9 and shorter bars indicating younger demographics in clusters like C1, highlighting the demographic diversity within different geographic segments.

6) How does the age of cars with different transmission types vary across different fuel types?

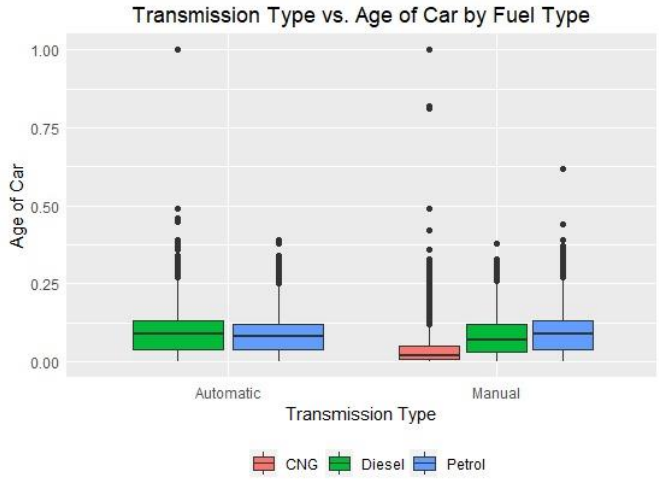


Figure 8: Transmission Type vs. Age of Car Colored by Fuel Type.

The boxplot shows the age of cars across different fuel types (CNG, Diesel, and Petrol) with different transmission types (Automatic and Manual). The median age is similar across all fuel types, but there are differences in age distribution within each category. The visualization suggests that age does not significantly differ between fuel types.

7) Are certain age groups of policyholders with specific transmission types more likely to file claims?

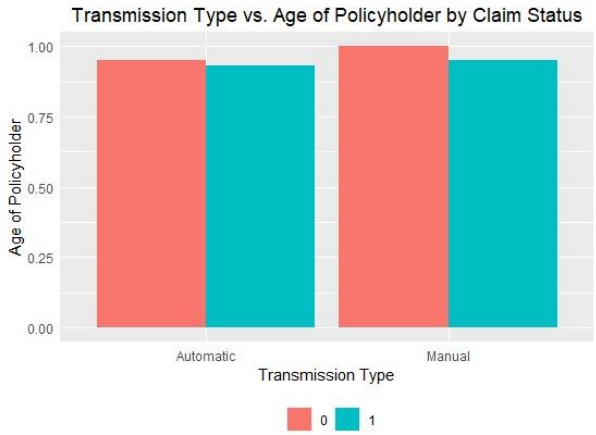


Figure 9: Transmission Type vs. Age of Policyholder Colored by Claim Status.

The chart compares policyholders' age and car transmission type (Automatic or Manual) against their claim status. It shows no significant difference in claim filing between age groups within transmission types. Both Automatic and Manual transmission types have similar heights for claim statuses. The chart does not show a clear trend that policyholders of a specific age group with a specific transmission type are more likely to file claims.

8) How does the NCAP rating of cars driven by different age groups of policyholders vary with transmission type?

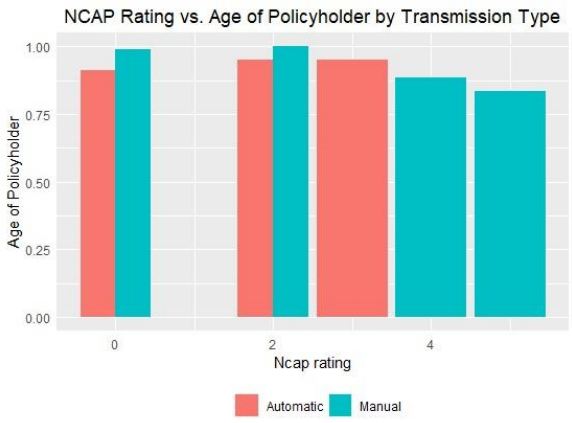


Figure 20: NCAP Rating vs. Age of Policyholder Colored by Transmission Type.

The chart shows a consistent relationship between the age of policyholders and NCAP ratings for both automatic and manual transmission vehicles. There is no significant variation in policyholder age with NCAP ratings, suggesting that policyholders drive cars with a range of safety ratings. Transmission type does not strongly correlate with safety rating, suggesting no distinct preference for higher or lower NCAP rated vehicles.

9) Does car height correlate with transmission and fuel type preferences?

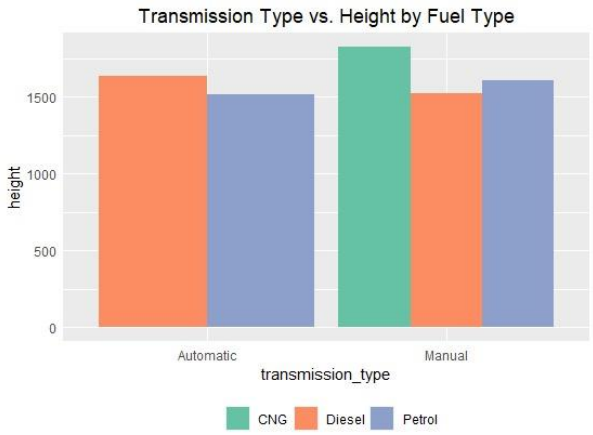


Figure 31: Transmission Type vs. Height Colored by Fuel Type.

The bar chart shows a correlation between car height and transmission and fuel type preferences. Diesel cars are the tallest, followed by CNG and Petrol vehicles. There is a slight variation in car height based on fuel type, but this is relatively small. Transmission type preference does not significantly influence car height, suggesting a correlation between car height and fuel type.

10) Is there a relationship between car width and the likelihood of a claim being filed, and does this differ by fuel type?

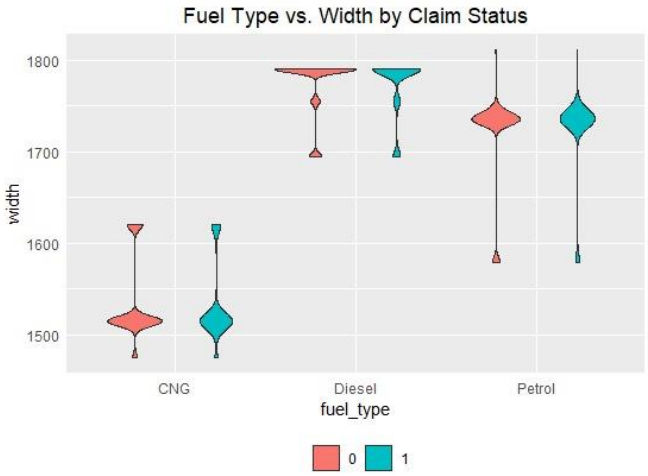


Figure 12: Fuel Type vs. Width Colored by Claim Status.

The violin plot shows the distribution of car widths for different fuel types and their corresponding claim statuses. The widths for cars with filed claims do not significantly differ from those without claims within the same fuel type category. However, there is a noticeable spread in the width of vehicles with filed claims in the Petrol category, suggesting that car width alone may not be a strong indicator of claim likelihood.

6. DATA PREPROCESSING

Data preprocessing is a critical step in the analytical modeling process. It involves transforming raw data into an understandable format for further processing and analysis. In this project, the following preprocessing steps were undertaken:

Handling Class Imbalance:
In predictive modeling, addressing class imbalance is crucial for accurate performance. The dataset revealed a significant skew in the 'is_claim' variable, impacting model generalization. To counter sthis, Synthetic Minority Over-sampling Technique (SMOTE) was applied, balancing the dataset by generating synthetic samples for the minority class. This ensures fair representation for both classes in training, leading to improved model robustness and unbiased predictive performance. The distribution of the target variable post-SMOTE shows an equal number of records for both classes, ensuring a fair ground for model training.

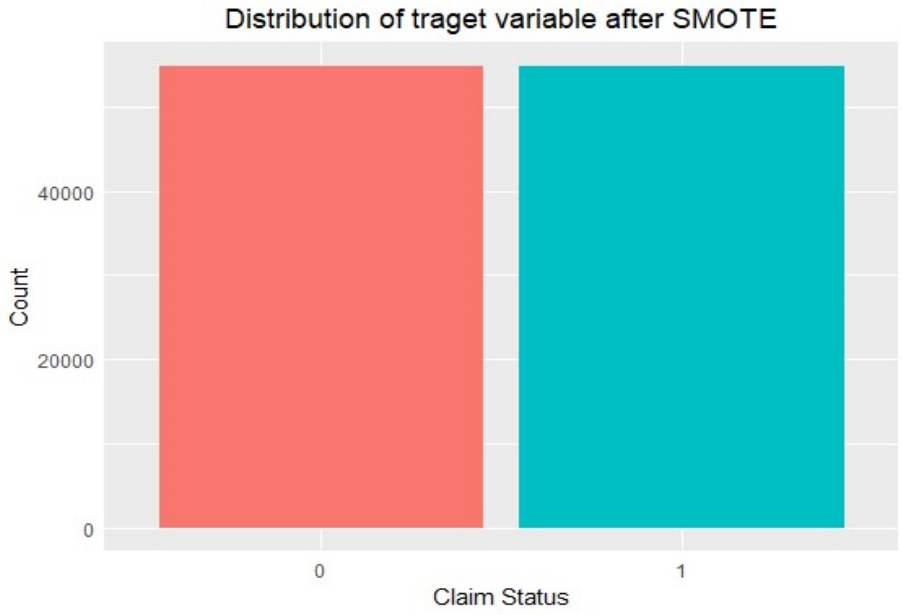


Figure 13: Distribution of target variable after SMOTE.

Separating Numerical and Categorical Variables:
The distinction between numerical and categorical variables, particularly when numerical data types might represent categories, is crucial for model accuracy. Incorrectly treating categorical variables as numerical can lead to erroneous conclusions and model predictions. Therefore, a careful separation was made where true numerical columns are treated for statistical modeling, and categorical columns, even those encoded as numbers, are handled appropriately, often through encoding techniques. The following are the numerical and categorical columns that are separated based on the required prediction.

Categorical columns

```
 [1] "make"                          "segment"                  "model"
 [4] "fuel_type"                     "max_torque"               "max_power"
 [7] "engine_type"                   "airbags"                  "is_esc"
[10] "is_adjustable_steering"        "is_tpms"                  "is_parking_sensors"
[13] "is_parking_camera"             "rear_brakes_type"         "displacement"
[16] "cylinder"                      "transmission_type"        "gear_box"
[19] "steering_type"                 "turning_radius"           "length"
[22] "width"                         "height"                   "gross_weight"
[25] "is_front_fog_lights"           "is_rear_window_wiper"     "is_rear_window_washer"
[28] "is_rear_window_defogger"       "is_brake_assist"          "is_power_door_locks"
[31] "is_central_locking"            "is_power_steering"        "is_driver_seat_height_adjustable"
[34] "is_day_night_rear_view_mirror" "is_ecw"                   "is_speed_alert"
[37] "ncap_rating"                   "is_claim"                 "area_cluster"
```

Numerical columns

```
[1] "age_of_car"          "age_of_policyholder" "population_density"
```

After checking the distribution, we observed that there are no outliers in our dataset, so we will not perform outlier removal.

Feature Selection and Reduction of Multicollinearity:

We employed label encoding for some columns, converted specific categorical variables into dummy variables, and further eliminated irrelevant features based on their correlation values to streamline the dataset for the intended prediction.

Data Splitting:

The processed dataset was split into a training set, which consisted of 75% of the data, and a test set, which made up the remaining 25%. This split was performed to validate the model's performance on unseen data, ensuring that the predictive models were robust and generalized well.

7. PREDICTIVE MODELING

Predictive modeling is integral to our Car Insurance Claim Prediction project. This section delves into the specifics of each model employed, their implementation, and the nuances of their predictive capabilities.

a) **Logistic Regression:** Utilized as the baseline model, Logistic Regression estimates the probability of a claim being filed using a logistic function. This model benefits from its simplicity, providing easily interpretable odds ratios for each predictor.To prevent overfitting, we applied L2 regularization, which penalizes large coefficients. The hyperparameters of the regularization term were optimized using cross-validation techniques.

b) Linear Discriminant Analysis: Chosen for datasets where predictors are assumed to have a Gaussian distribution, LDA seeks to reduce the feature space by projecting it onto a lower-dimensional space that maximizes class separation. Prior to applying LDA, we verified the assumption of normality and equal variance across the classes. We also tuned the model to determine the most discriminative features.

c) **Random Forest Regression:** This ensemble method aggregates the decisions from multiple decision trees to improve the model's predictive accuracy and robustness. Each tree in the Random Forest is built on a bootstrap sample of the data, and the final prediction is made by averaging the predictions (regression) or taking a majority vote (classification) of the trees. We conducted a grid search to find the optimal combination of hyperparameters such as the number of trees, the maximum depth of the trees, and the minimum number of samples required to be at a leaf node.

Model Training:

The training phase was methodically executed, ensuring that each model was exposed to a wide variety of data scenarios through k-fold cross-validation. The test set, a separate data partition, provided an unbiased platform to assess each model's performance.

Evaluation Metrics:

We employed accuracy, precision, recall, F1 score, and the ROC AUC as our performance metrics. These metrics allowed us to assess not only the correctness of the models' predictions but also their ability to manage the trade-offs between true positive and false positive rates.

Performance Summary Table:

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	53.43%	53.45%	53.19%	53.32%	0.555
Linear Discriminant Analysis	53.59%	53.61%	53.29%	53.45%	0.555
Random Forest Regression	69.14%	70.82%	65.11%	67.85%	0.755

8. FINDINGS

In the performance summary, the Logistic Regression and Linear Discriminant Analysis models exhibit similar metrics around 53%, while the Random Forest Regression outperforms with an accuracy of 69.14% and superior precision, recall, F1 score, and ROC AUC (70.82%, 65.11%, 67.85%, and 0.755, respectively). This suggests that the Random Forest model is more effective for the project's prediction task, warranting further consideration in the report.

9. CONCLUSION

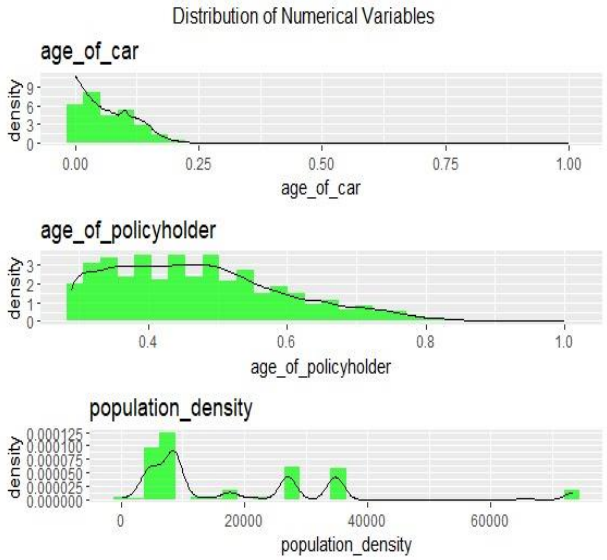
The project successfully developed a robust predictive model for forecasting insurance claims over a six-month period. Rigorous data preprocessing, including SMOTE for class balance and careful encoding of categorical variables, laid a solid foundation. Evaluating Logistic Regression, Linear Discriminant Analysis, and Random Forest models revealed Random Forest's superiority with 69.14% accuracy and an ROC AUC of 0.755. This emphasizes the efficacy of ensemble techniques in insurance predictive modeling, with Random Forest recommended for deployment due to its adeptness in navigating claim prediction nuances.

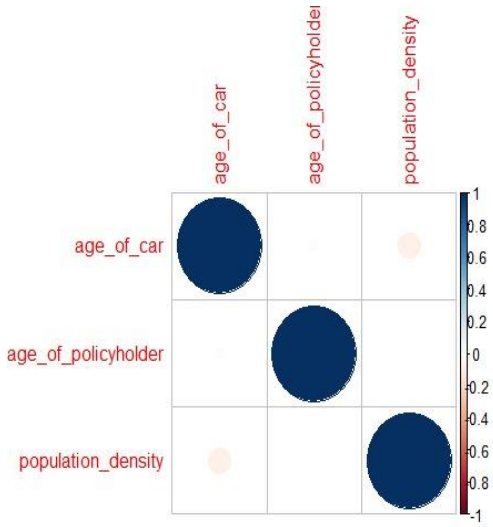
10. FUTURE WORKS

Future work should focus on refining the Random Forest model further through more nuanced hyperparameter tuning and exploring additional ensemble techniques. Moreover, efforts to interpret the model's decision-making process could provide actionable insights for business strategy. This project underscores the potential of machine learning in enhancing decision-making in the insurance industry, setting a precedent for future analytical endeavors in this domain.

APPENDIX:

Distribution of numerical variables and correlation:





Evaluation metrics of Logistic Regression:

```
[1] "Accuracy: 0.534388447232149"
> print(paste("Precision:", precision_result$.estimate))
[1] "Precision: 0.534559847540863"
> print(paste("Recall:", recall_result$.estimate))
[1] "Recall: 0.53190868645613"
> print(paste("F1 Score:", f1_result$.estimate))
[1] "F1 Score: 0.533230971704321"
```

Evaluation metrics of Linear Discriminant Analysis:

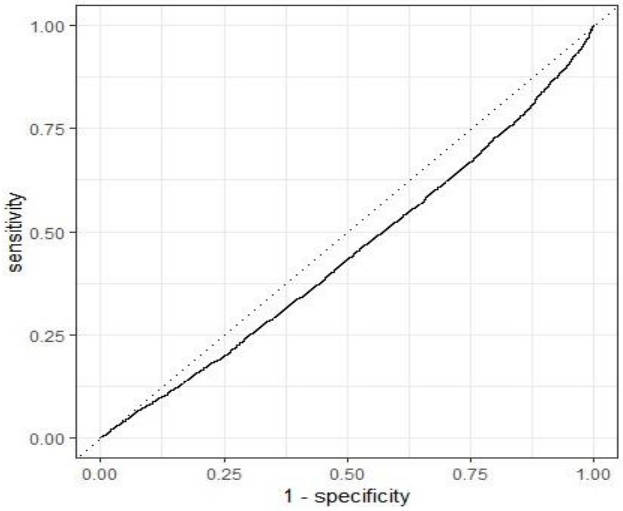
```
[1] "Accuracy: 0.535920064182044"
> print(paste("Precision:", precision_result$.estimate))
[1] "Precision: 0.536136180203977"
> print(paste("Recall:", recall_result$.estimate))
[1] "Recall: 0.532929764422726"
> print(paste("F1 Score:", f1_result$.estimate))
[1] "F1 Score: 0.534528163862472"
```

Evaluation metrics of Random Forest Regression:

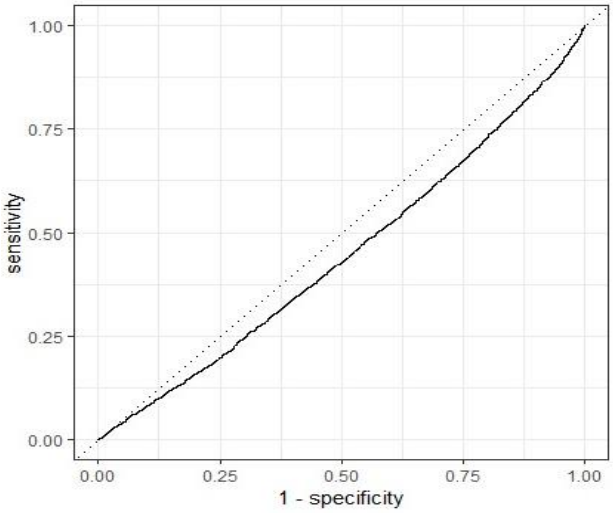
```
[1] "Accuracy: 0.691452118736781"
> print(paste("Precision:", precision_result$.estimate))
[1] "Precision: 0.708234174202761"
> print(paste("Recall:", recall_result$.estimate))
[1] "Recall: 0.651156006126468"
> print(paste("F1 Score:", f1_result$.estimate))
[1] "F1 Score: 0.678496789147699"
```

ROC curves:

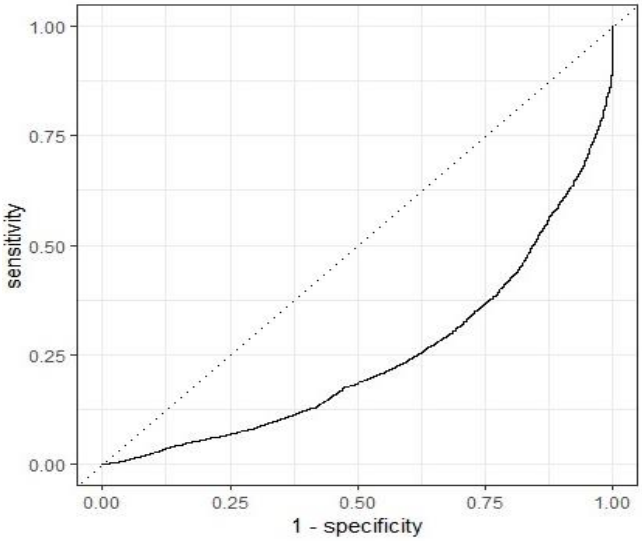
Logistic Regression Model



Linear Discriminant Analysis



Random Forest Regression



References

Chugh, V. (n.d.). *Logistic Regression in R Tutorial*. Retrieved from Datacamp: https://www.datacamp.com/tutorial/logistic-regression-R

Najnin, I. (n.d.). *Car Insurance Claim Prediction*. Retrieved from Kaggle: https://www.kaggle.com/datasets/ifteshanajnin/carinsuranceclaimprediction-classification/

Random Forests. (n.d.). Retrieved from Datacamp: https://campus.datacamp.com/courses/supervised-learning-in-r-regression/tree-based-methods

Zach. (2020, October 30). *Linear Discriminant Analysis in R (Step-by-Step)*. Retrieved from Statology: https://www.statology.org/linear-discriminant-analysis-in-r/