# Introduction

# In this report, we aim to develop a predictive model for vehicle claims using various features. These features include car details such as maker, model, adv year, adv month, color, registration year, body type, miles driven, engine size, gearbox, fuel type, price, seat number, door number, issue, issue id, adv day, breakdown date, repair complexity, repair cost, repair hours, category anomaly, repair date.

# The objective is to build an accurate model that predicts the likelihood of a vehicle claim. We will evaluate the model's performance using precision, recall, accuracy, and the F1 score.

# Preliminary analysis suggests that key features, such as category anomaly, the color gelb, repair hours, seat number, number of doors, and the model Focus.

# Exploratory Data Analysis

Target Variable: Vehicle Claims

The target variable in the model is whether a vehicle has a claim or not. It is a binary classification (claim/no claim). Understanding how the target variable relates to key features helps to identify which attributes have the most predictive power.

Key Variables in the Model:

Several key features were selected for model building, based on their importance and relationship to vehicle claims. Below is an overview of these key variables:

1. Category Anomaly:

* Description: This feature indicates whether there are anomalies in the car's category

2. Repair Hours:

* Description: This variable records the time spent on repairing the vehicle.

3. Seat Num:

* Description: Represents the number of seats in the car.

4. Door Num:

* Description: Indicates the number of doors in the vehicle.

5. Model Focus:

* Description: A binary variable that indicates whether the car model is a "Focus."



The key features in our analysis were carefully selected for their strong relationship with the likelihood of a claim, as highlighted in the heatmap. Here's how each key feature impacts the predictions:

* Category Anomaly: This feature shows the strongest correlation with claims. Vehicles flagged with a category anomaly often exhibit structural or design issues, which increase their chances of getting into accidents or having breakdowns, thereby raising the likelihood of a claim.
* Repair Hours: As the second most correlated feature, repair hours provide crucial insight into the severity of issues. Vehicles requiring longer repair times often face significant problems, making this feature a strong predictor of claim rates.
* Seat Number and Door Number: These features rank third and fourth in terms of correlation. Larger vehicles, characterized by higher seat and door counts, tend to have more claims. This could be attributed to greater usage, varying driving conditions, or the inherent complexity of larger designs. These features help capture patterns related to vehicle size and function.
* Model Focus: Although this feature has the lowest correlation with claims, it plays a meaningful role in improving the F1 score. Specific models, such as the Ford Focus, are associated with a higher frequency of claims due to known issues or historical trends, making this feature valuable for refining predictions.

In summary, our analysis shows that vehicles with certain characteristics—such as longer repair times, specific models, or unique categories—are more likely to have claims. These features help our model make better predictions.

# Model Evaluation

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Features | Avg accuracy | SD of accuracy | Avg precision | SD precision | Avg recall | SD recall | Avg f1 | SD f1 |
| Model 1 | "category\_anomaly", "Color\_Gelb", "Repair\_Hours\_Binned\_Extreme", "Seat\_num", "repair\_cost" | 0.9006 | 0.0111 | 0.9035 | 0.0095 | 0.9006 | 0.0111 | 0.8902 | 0.0141 |
| Model 2 | "category\_anomaly", "repair\_hours", "Seat\_num", "Door\_num", "Model\_Focus" | 0.9440 | 0.0094 | 0.9473 | 0.0085 | 0.9440 | 0.0094 | 0.9404 | 0.0105 |
| Model 3 | *"category\_anomaly", "repair\_hours", "Seat\_num"* | 0.9378 | 0.0099 | 0.9380 | 0.0092 | 0.9378 | 0.0099 | 0.9346 | 0.0120 |

# Conclusion

Model selected as the best model is model 2 that has 1 binned feature and dummying for all the categorical features. The reasons why it is chosen are:

1. Higher Performance Metrics  
   Model 2 outperforms the other models across the key metrics, getting an average accuracy of 0.9440, precision of 0.9473, recall of 0.9440, and an F1 score of 0.9404. These metrics indicate that Model 2 is both very accurate and well-balanced in terms of precision and recall.
2. Simplified Feature Engineering  
   Model 2 has a simplified approach to feature engineering, using fewer bins and avoiding complex transformations. It focuses on critical features such as "category\_anomaly," "repair\_hours," "Seat\_num," "Door\_num," and "Model\_Focus," which capture the essential predictors without unnecessary complications.
3. Streamlined Preprocessing  
   The preprocessing for Model 2 is minimal, which makes interpretability easier and reduces overhead. By avoiding overly complex scaling or transformations, it ensures the model is straightforward to implement and understand.

Conclusion:  
Model 2 is chosen because it achieves a good balance between performance, simplicity, and interpretability. Its high accuracy and streamlined design make it suitable for the task.