

Executive Summary: Predicting Car Insurance Claims

On the Road Car Insurance requested assistance in developing a model to predict whether a customer will file a claim during their policy period. The goal was to identify the most influential feature for the model to aid in deployment, as the company has limited infrastructure for monitoring complex models.

The provided dataset, `car_insurance.csv`, includes various customer-related features and a target variable, `outcome`, which indicates if a claim was made (1) or not (0). The key features in the dataset are:

- **age, gender, driving_experience, education, income, credit_score, vehicle_ownership, vehicle_year, married, children, annual_mileage, vehicle_type, speeding_violations, dui, past_accidents.**

To solve the problem, we used logistic regression models, where each feature was tested individually to assess its predictive power regarding the likelihood of making a claim. The following steps were taken:

1. **Data Preprocessing:**
 - We loaded and cleaned the dataset, removing unnecessary columns (e.g., `id`) and focusing on features that would predict the target variable (`outcome`).
2. **Model Development:**
 - A logistic regression model was built for each feature, with the `outcome` as the dependent variable. The `Logit` function from the `statsmodels` library was used to fit the models and evaluate performance.
3. **Accuracy Calculation:**
 - The performance of each model was assessed by calculating accuracy using the confusion matrix (true positives, true negatives, false positives, false negatives).
4. **Best Performing Feature:**
 - The model with the highest accuracy identified **driving_experience** as the most predictive feature, with an accuracy of 77.71%.

Conclusion:

The analysis revealed that **driving experience** is the strongest predictor of car insurance claims, achieving an accuracy of 77.71%. This allows On the Road Car Insurance to deploy a simple and effective model using this feature to predict customer claims.