SANLAM INSURANCE

Accident Report for Sanlam Insurance

By Charles Ndegwa

OVERVIEW

The goal of the report is to enhance the underwriting process by leveraging data analytics to gain insights into the factors contributing to severe crashes

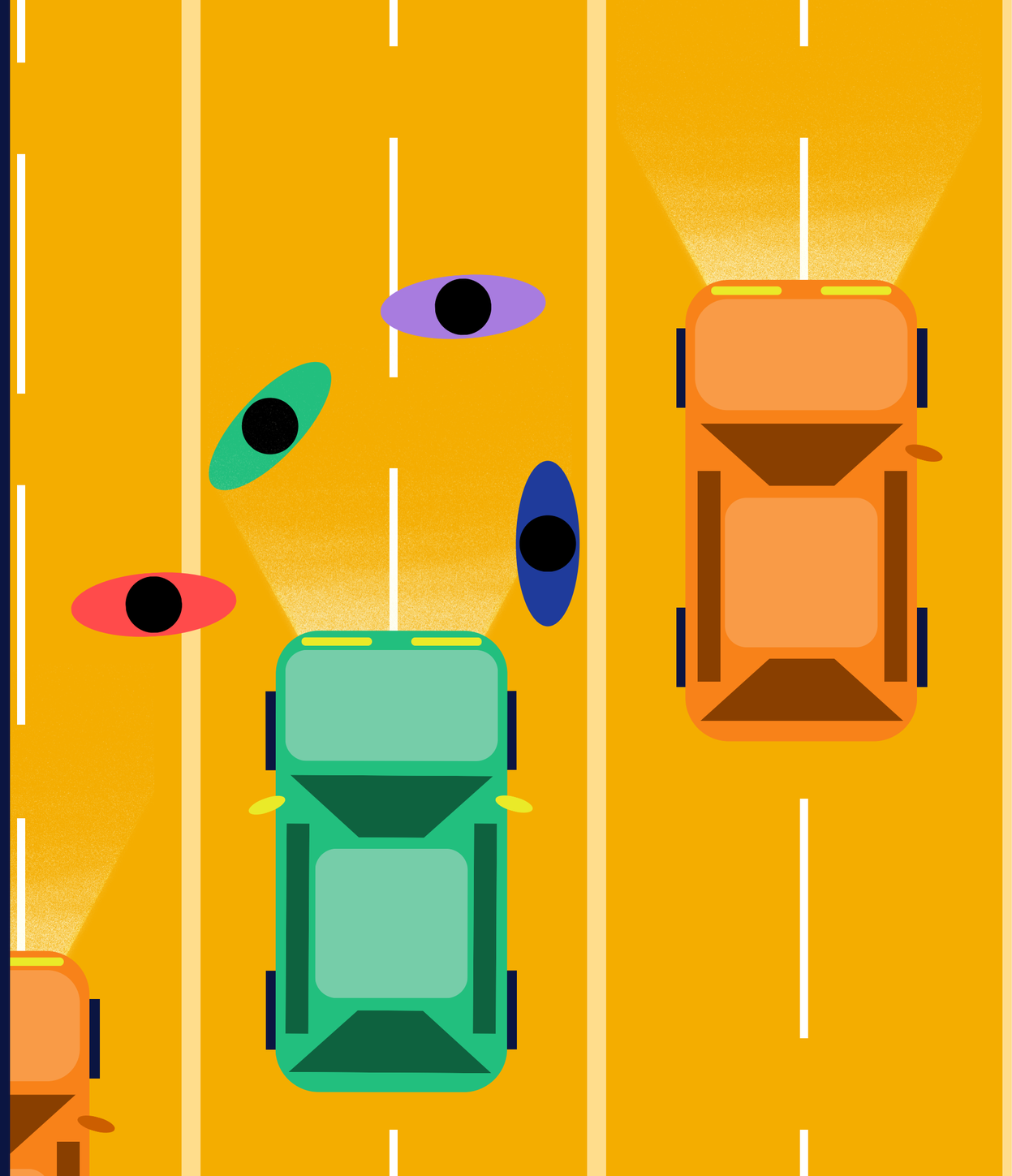
1. **INJURY AND / OR TOW DUE TO CRASH**
2. **NO INJURY / DRIVE AWAY**

DATA

The project wasa clasiffier model where data from the Chicago Data Portal - (https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/about_data)

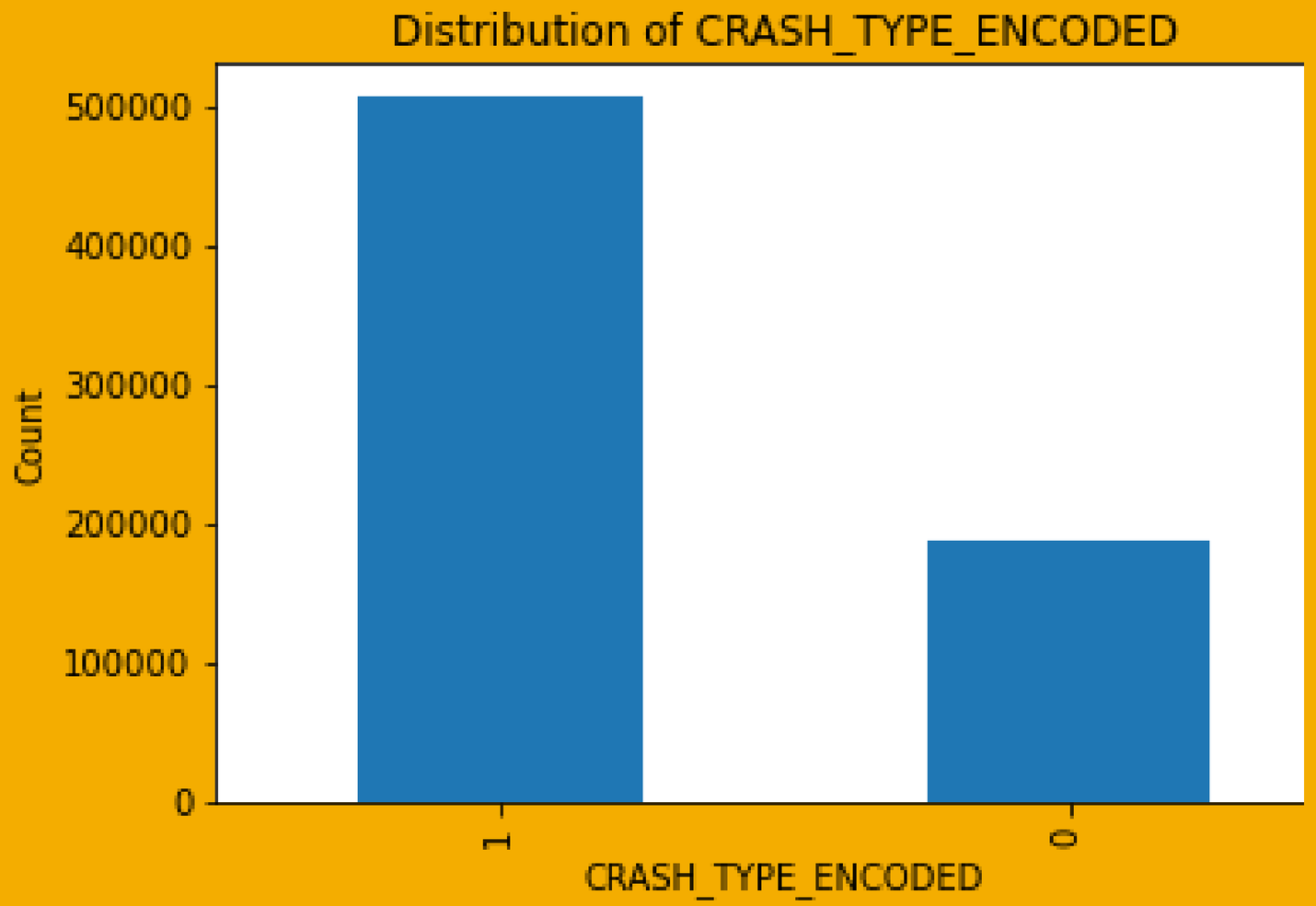
METHODOLOGY

Encode target Variable
Select Independent features
Handle missing values
Univariate Analysis
Bivariate Analysis
Modelling
Evaluation
Conclusions
Recommendations
Next steps



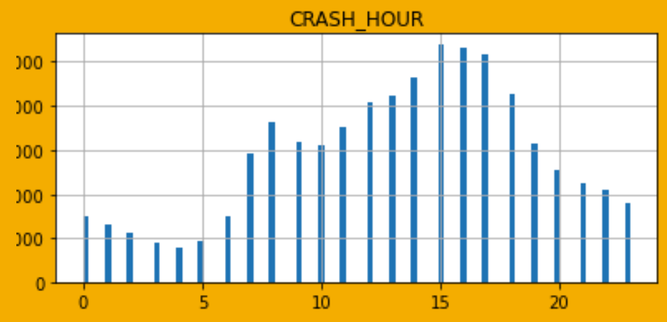
TARGET VARIABLE

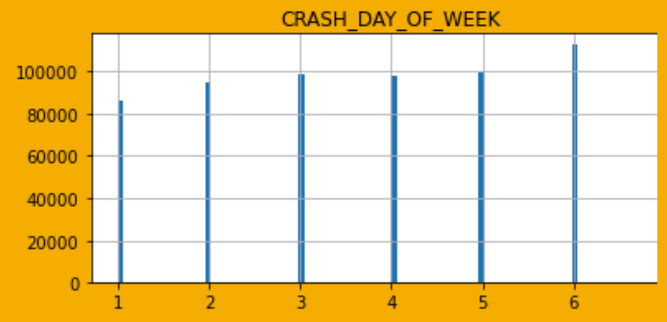
The precesence of imbalnce notice from target variable

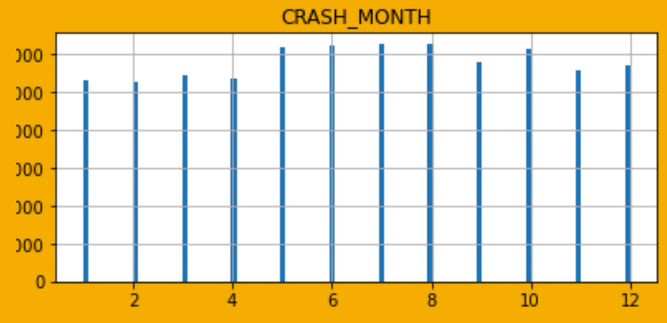


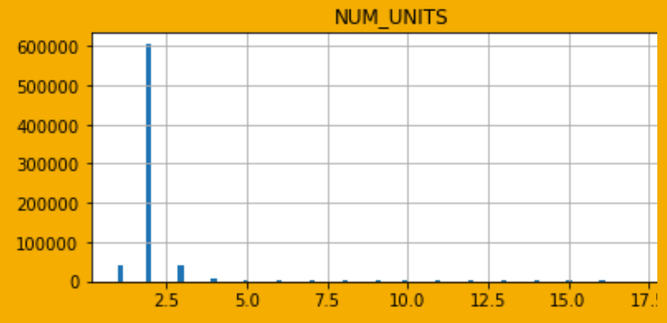
NUMERICAL FEATURES

Distribution for the numerical features shows hour, day and week has normal distribution while num_units and street no have a skewed distribution



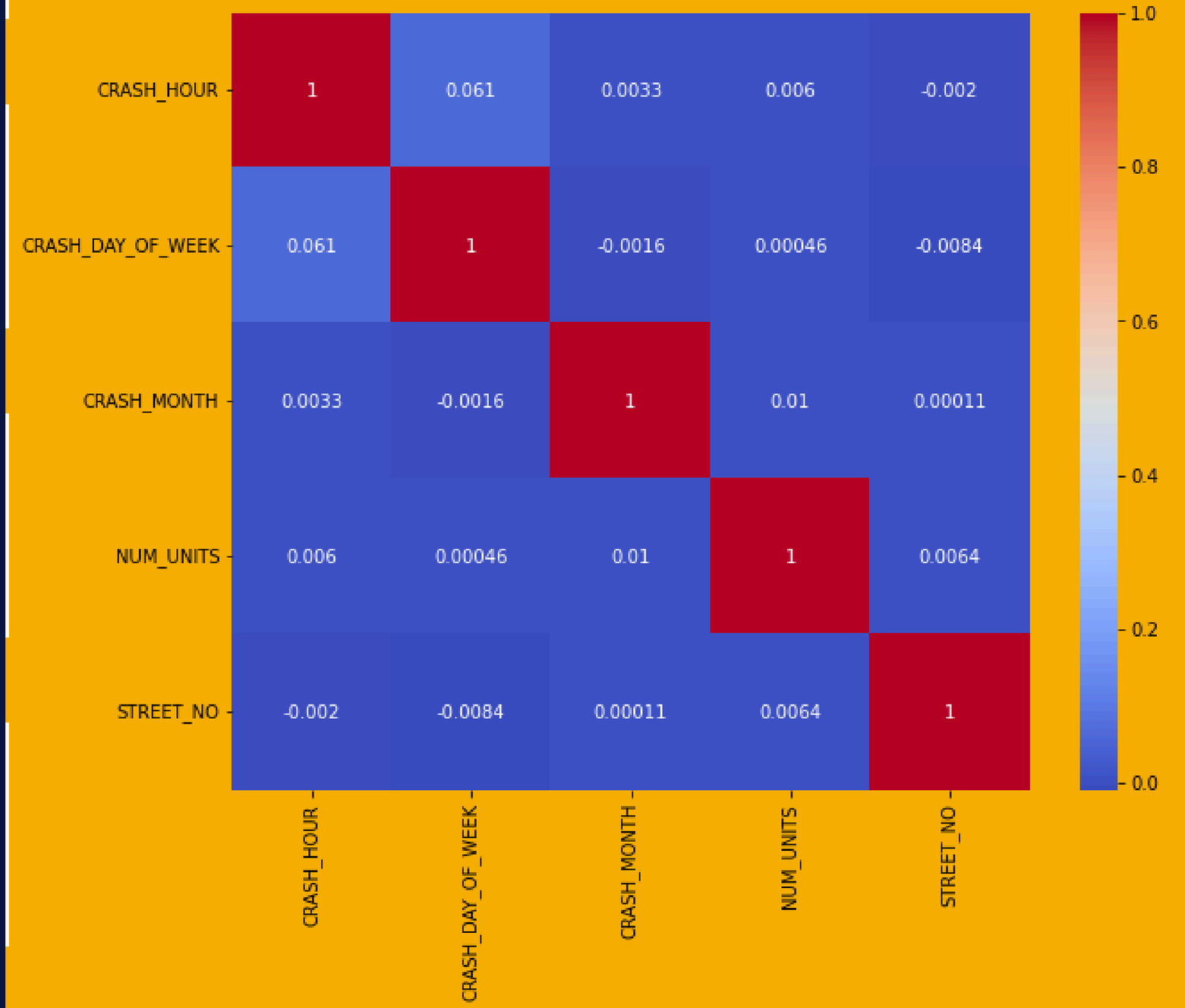






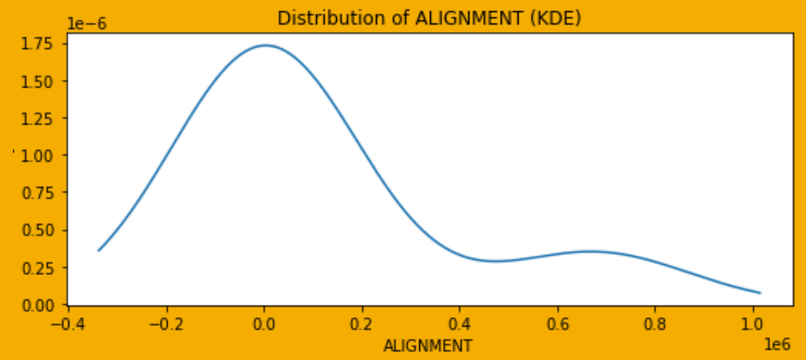
NUMERICAL FEATURES

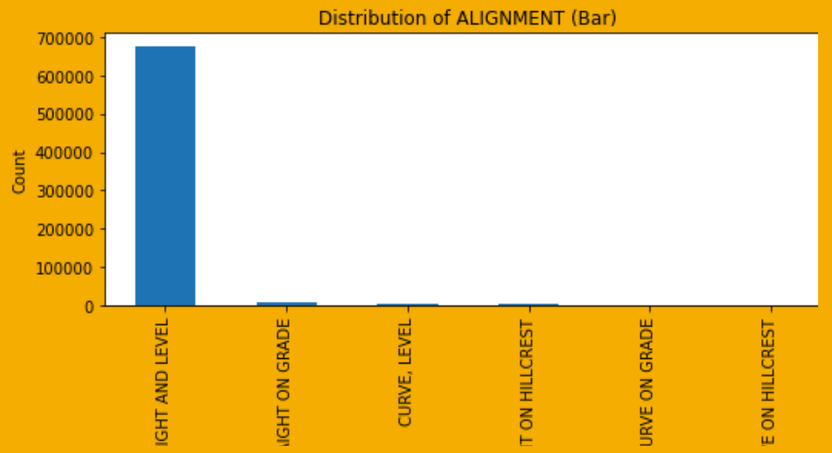
low to non existent correlation between the numerical columns

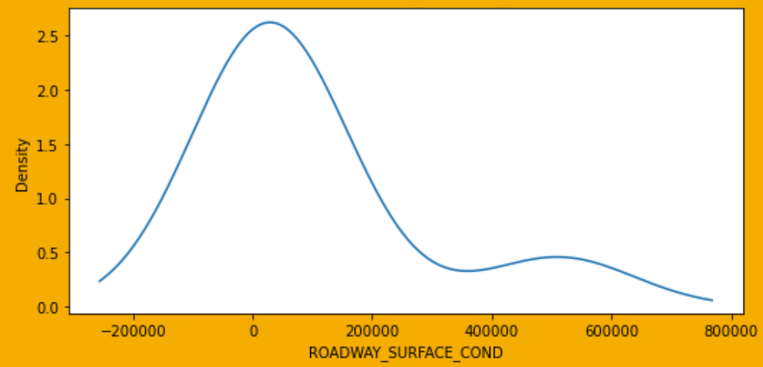


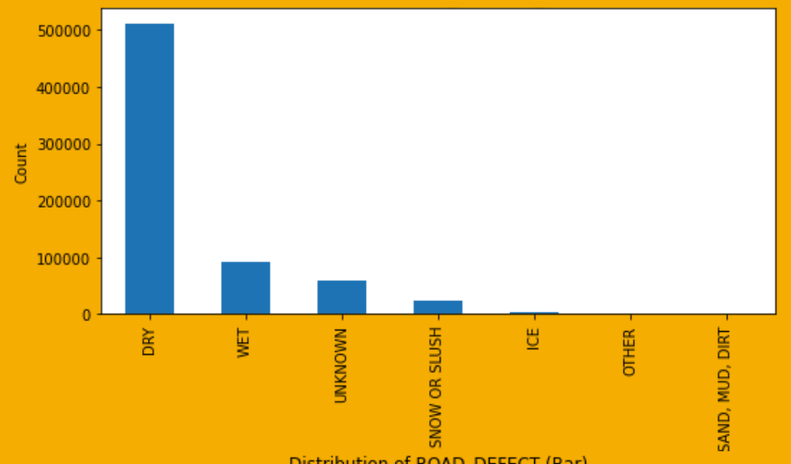
CATEGORICAL FEATURES

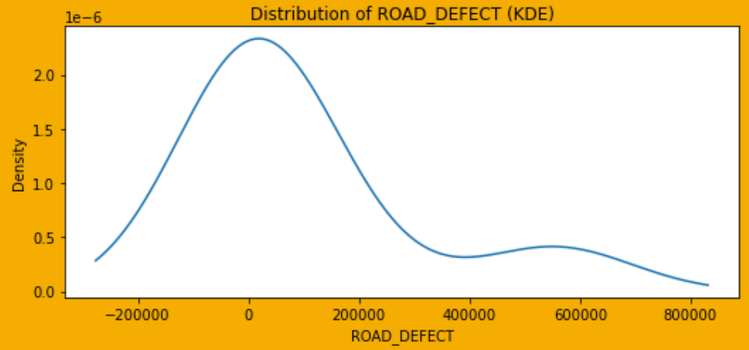
Most Categorical features skewed .

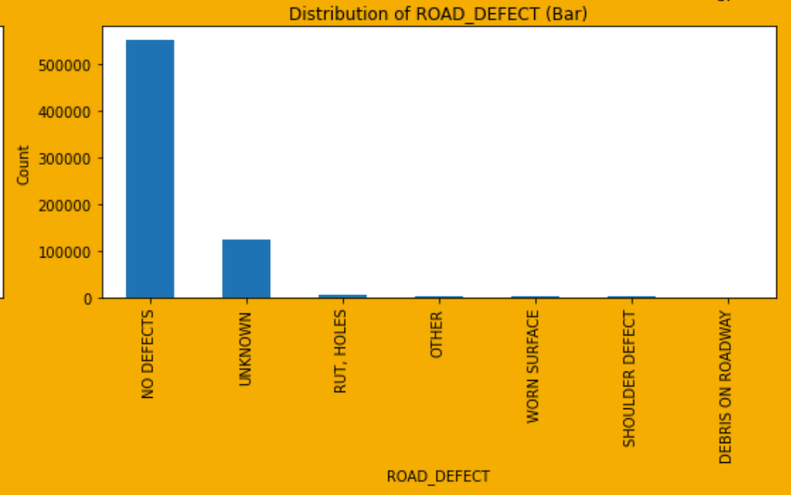






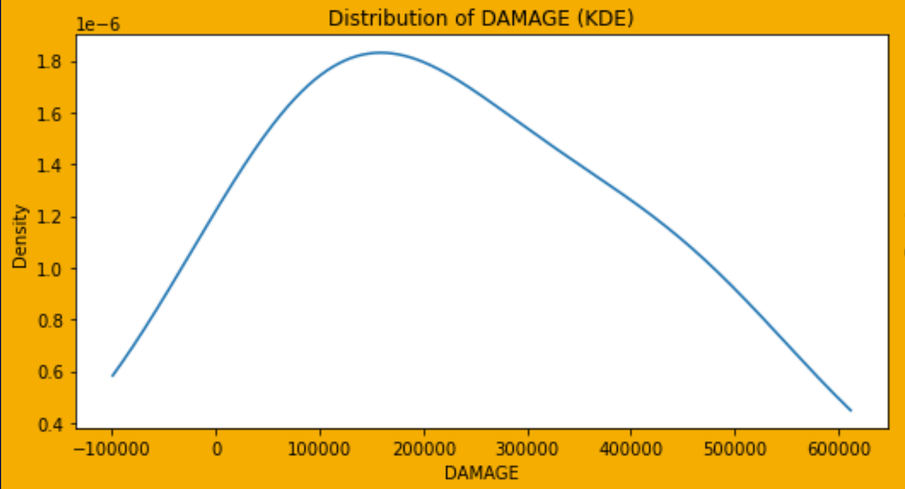


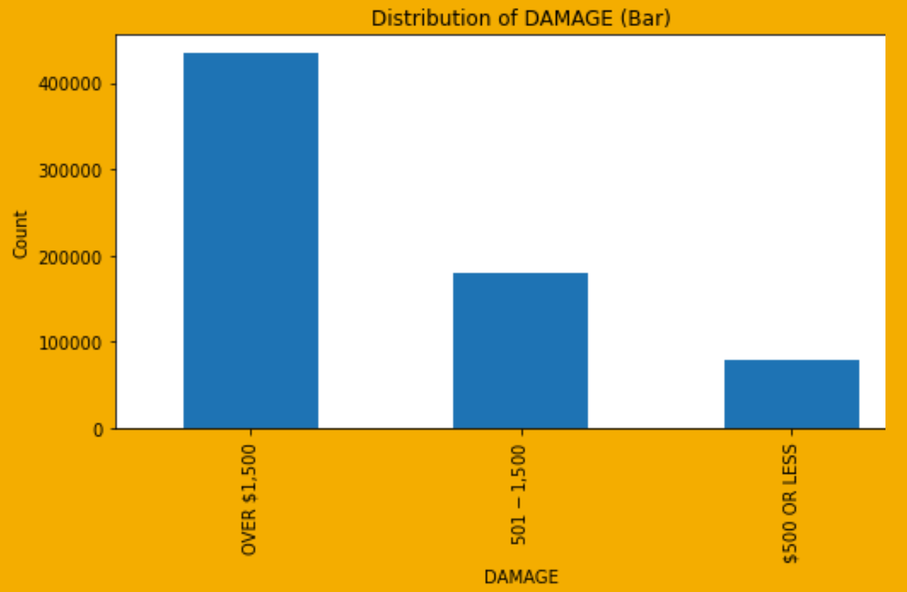


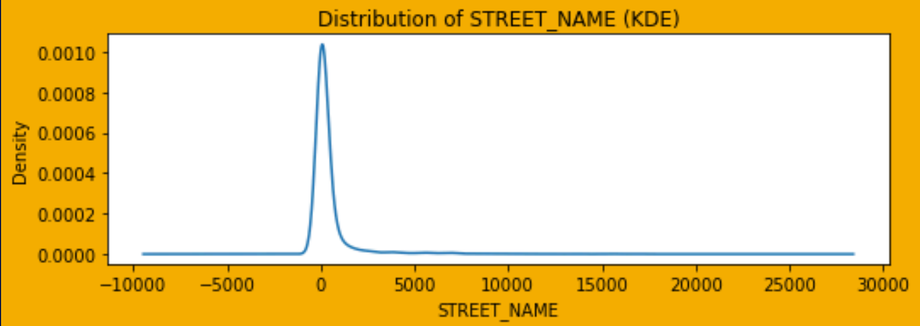


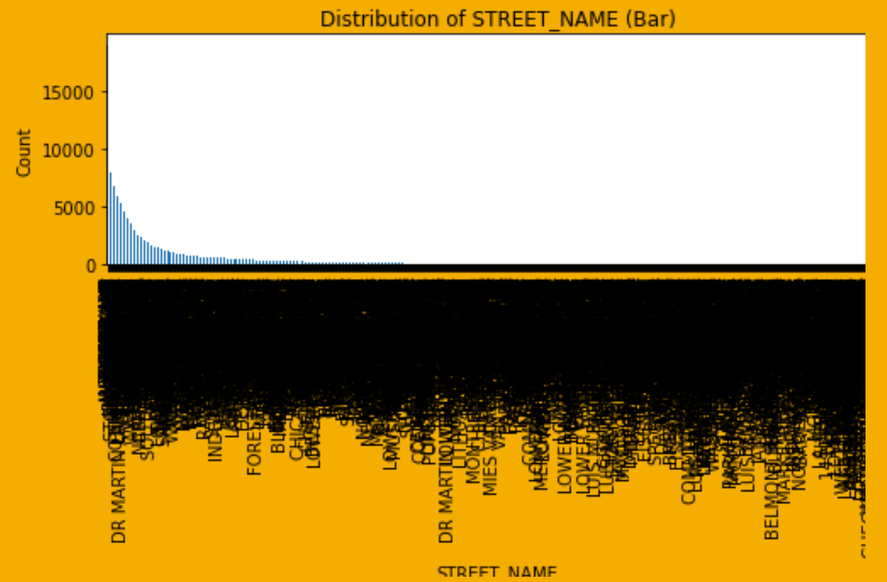
CATEGORICAL FEATURES

continued....



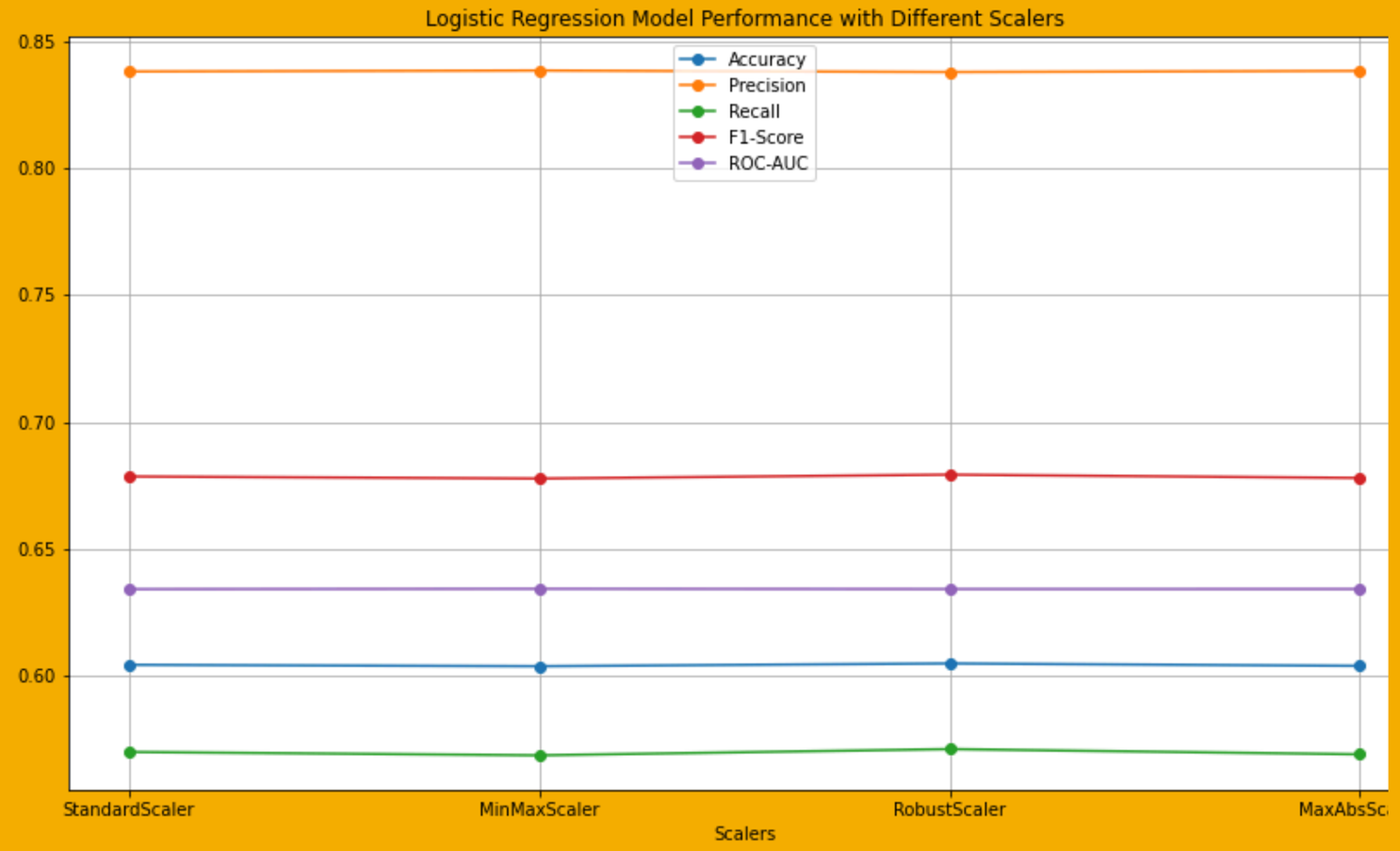






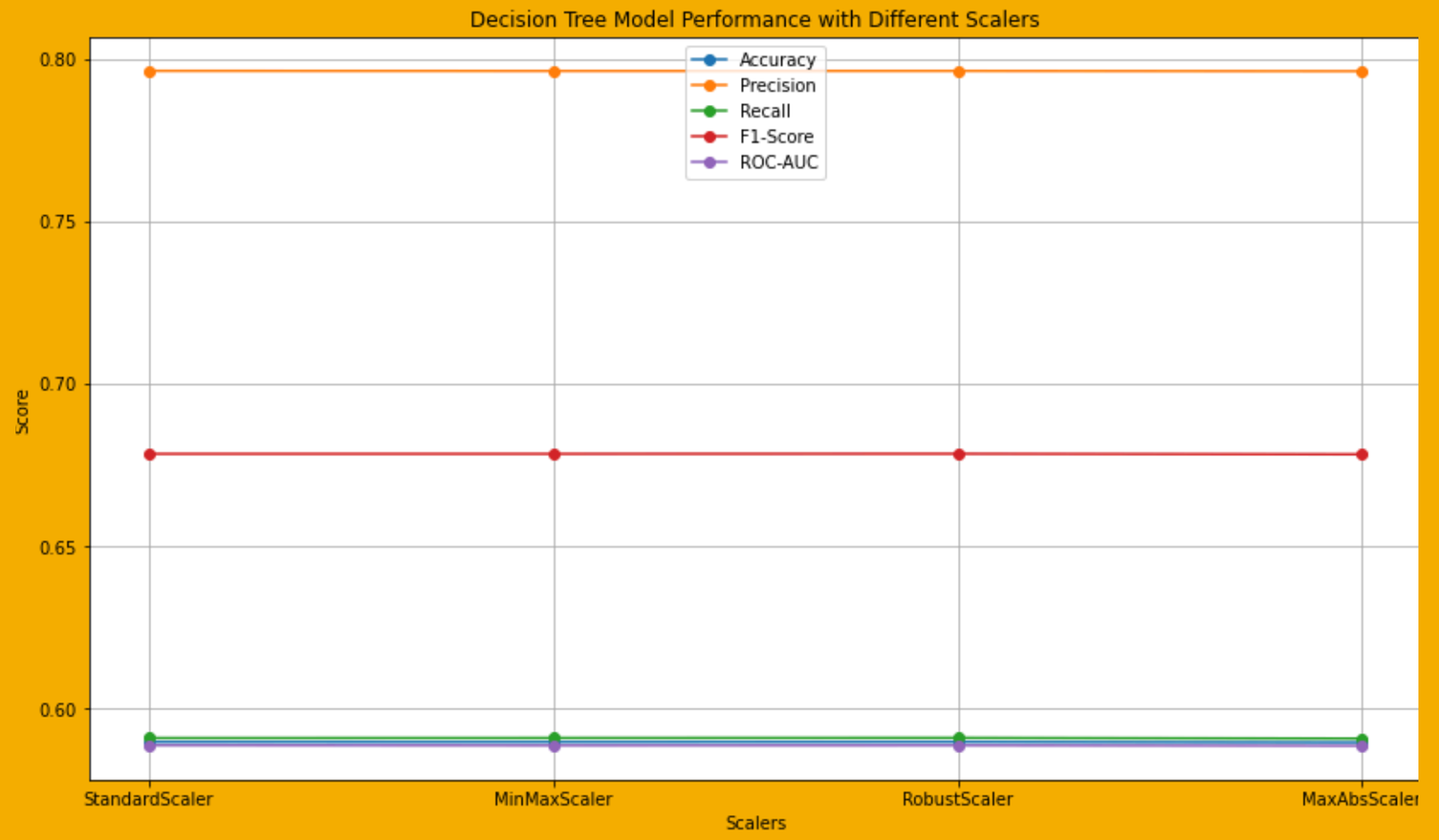
MODEL 1

Baseline model - Logistical regression with parameter random_state = 42
The different scalers were the criteria for iteration through the model



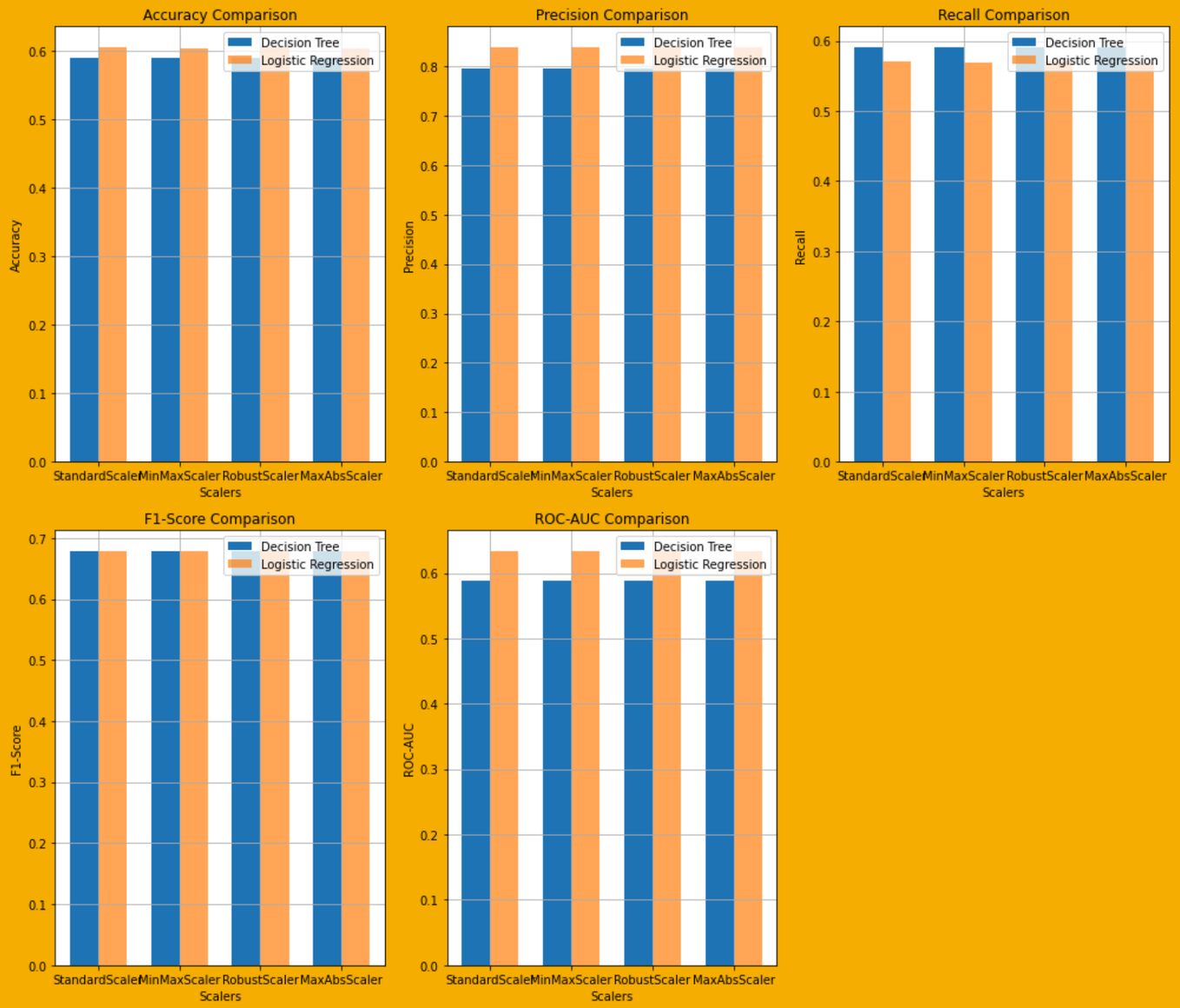
MODEL 2

Decision Tree classifier with parameter random_state=42



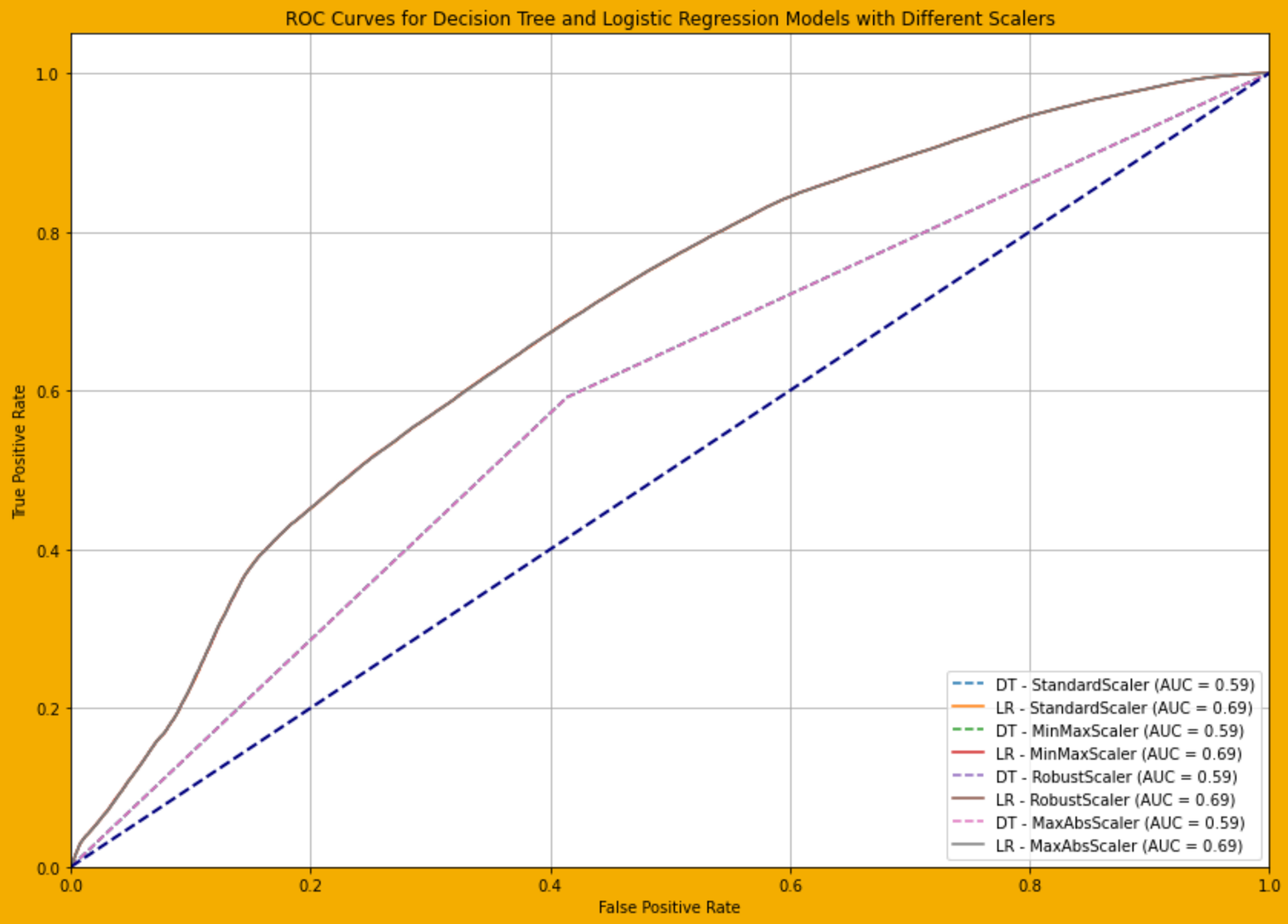
MODELLING

the bar charts illustrastes the difference in the models under different scalers



EVALUATION

Both models in the classification were able to be above 50% of the AUC but due to the low recall rate the models were not optimal for production



RECOMMENDATIONS

Feature Engineering: Explore creating new features or transforming existing ones to capture more relevant information.

Hyperparameter Tuning: Fine-tune the hyperparameters of both models to optimize their performance.

Ensemble Methods: Combine multiple models (e.g., using random forests or gradient boosting) to potentially improve overall performance and reduce overfitting.

Consider Other Models: Experiment with other machine learning algorithms that might be better suited to your specific problem.

Domain Expertise: Leverage insights from insurance experts to identify additional factors that might influence crash severity.

Cross-Validation: Use cross-validation techniques to ensure that the model performance observed is generalizable and not due to specificities in the train-test split. This will give a better estimate of model performance across different data subsets.

NEXT STEPS

Data Quality: Ensure the quality and completeness of your data to avoid biases in your analysis. Using other datasets analyze features that would increase model efficiency

Cost-Sensitive Learning: Given the high cost of missing severe crashes, explore cost-sensitive learning approaches where the model penalizes false negatives more heavily. This could help improve the recall for severe crashes, which is critical for the business problem.

PRESENTED BY:

CHARLES NDEGWA