Vehicle Insurance Fraud Detection

Group 1:
Anushree Chowdhary
James Le
Sophia Nguyen
Swarnima Sharma

Understanding the context

Business Question: *How Can We Effectively Detect Fraudulent Insurance Claims?*

Content:

•**Problem**: Fraudulent claims increase operational costs for insurers.

•**Goal**: Develop a machine learning model to automate fraud detection, improving efficiency and reducing losses.

Background and Context:

- **Background**: Fraud in Insurance include staged accidents, phantom passengers, and false injury claims. Globally, billions are lost annually to insurance fraud.
- **Dataset Details**: Attributes include vehicle details, policy information, and accident specifics.
- **Target**: FraudFound_P (indicates whether a claim is fraudulent).
- **Project Objectives**: Build a predictive model to classify insurance claims as fraudulent or non-fraudulent. Address the imbalance in the dataset (fraud cases are rare). Provide actionable insights to insurance companies.

Pre-Processing

Feature Selection

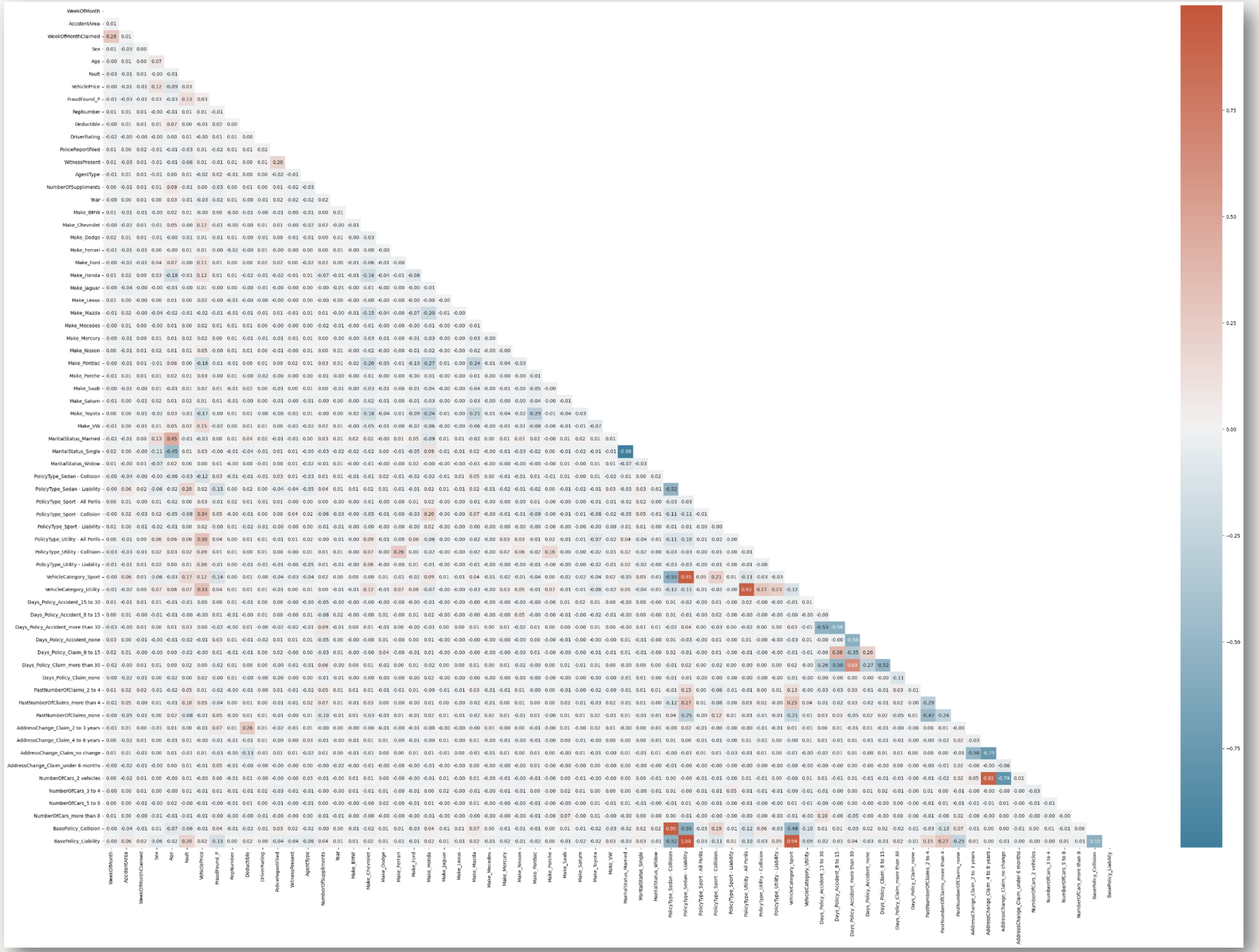
- Dropped irrelevant or overly complex features (e.g., make and policy number)
- Addressed multicollinearity: no strongly correlated variables detected and no high VIFs

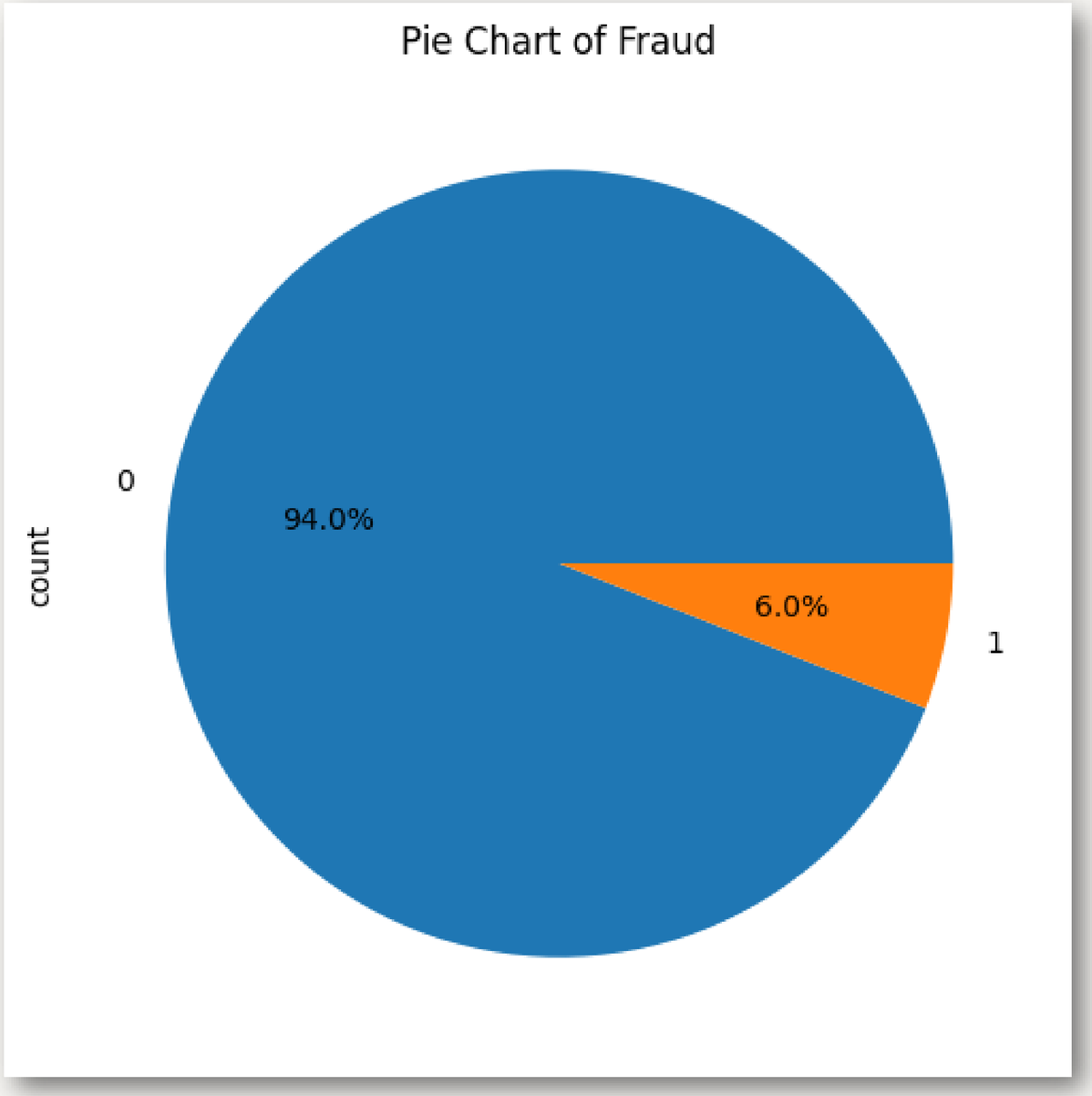
Data Transformation

- Converted categorical variables to binary/dummy variables
- Scaled numerical variables to ensure consistent contributions across models.

Evaluation Preparation

- Addressed model assumptions (e.g., independent predictors).
- Prepared data for model performance evaluation using metrics such as accuracy, recall, and F1 score.


WeekOfMonth
AccidentArea
WeekOfMonthClaimed
Sex
Age
Fault
VehiclePrice
FraudFound_P
RepNumber
Deductible
DriverRating
PoliceReportFiled
WitnessPresent
AgentType
NumberOfSuppliments
Year
Make_BMW
Make_Chevrolet
Make_Dodge
Make_Ferrari
Make_Ford
Make_Honda
Make_Jaguar
Make_Lexus
Make_Mazda
Make_Mercedes
Make_Mercury
Make_Nissan
Make_Pontiac
Make_Porche
Make_Saab
Make_Saturn
Make_Toyota
Make_VW
MaritalStatus_Married
MaritalStatus_Single
MaritalStatus_Widow
PolicyType_Sedan - Collision
PolicyType_Sedan - Liability
PolicyType_Sport - All Perils
PolicyType_Sport - Collision
PolicyType_Sport - Liability
PolicyType_Utility - All Perils
PolicyType_Utility - Collision
PolicyType_Utility - Liability
VehicleCategory_Sport
VehicleCategory_Utility
Days_Policy_Accident_15 to 30
Days_Policy_Accident_8 to 15
Days_Policy_Accident_more than 30
Days_Policy_Accident_none
Days_Policy_Claim_8 to 15
Days_Policy_Claim_more than 30
Days_Policy_Claim_none
PastNumberOfClaims_2 to 4
PastNumberOfClaims_more than 4
PastNumberOfClaims_none
AddressChange_Claim_2 to 3 years
AddressChange_Claim_4 to 8 years
AddressChange_Claim_no change
AddressChange_Claim_under 6 months
NumberOfCars_2 vehicles
NumberOfCars_3 to 4
NumberOfCars_5 to 8
NumberOfCars_more than 8
BasePolicy_Collision
BasePolicy_Liability
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Pie Chart of Fraud
count
0
94.0%
6.0%
1

Exploratory Analysis

Numerical Variable distribution

- Age slightly skewed right – peak around late 20s to early 30s

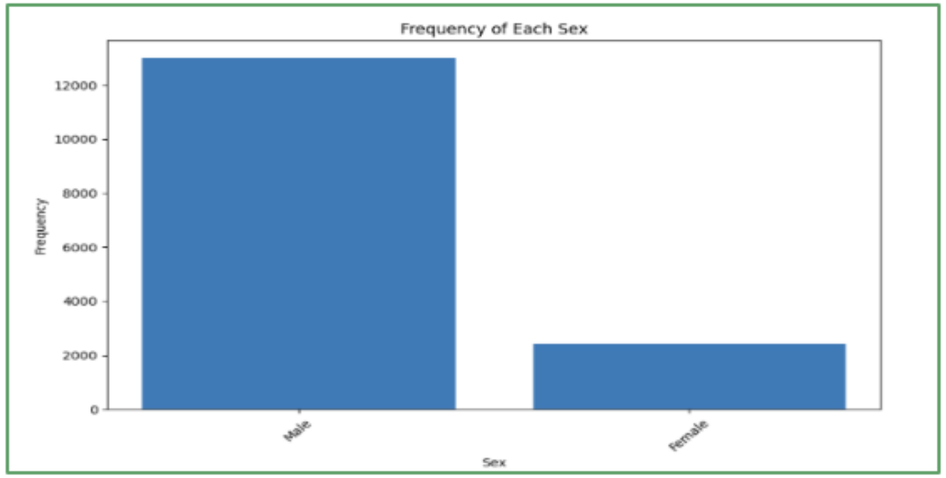
Univariate Analysis

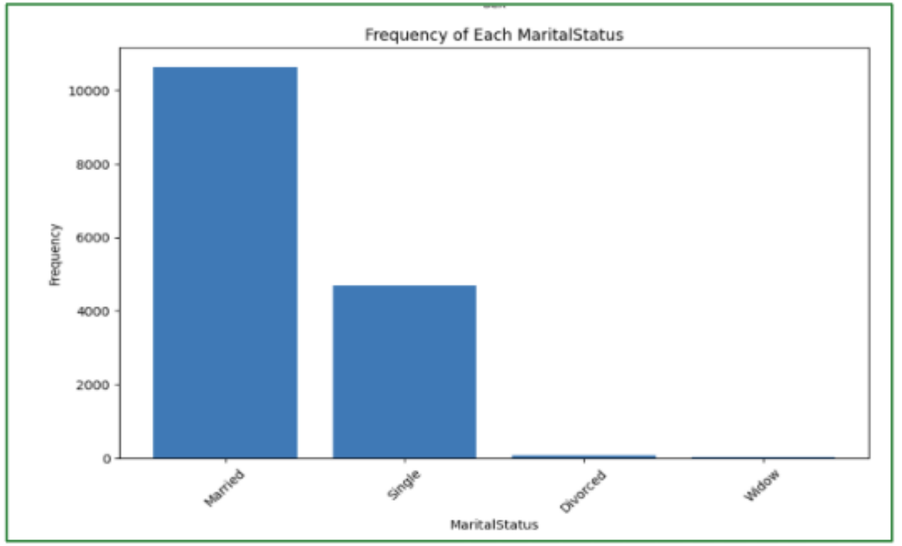
- Claims - more male policyholders, older vehicles, urban areas, and certain makes.
- Fraud patterns show seasonality

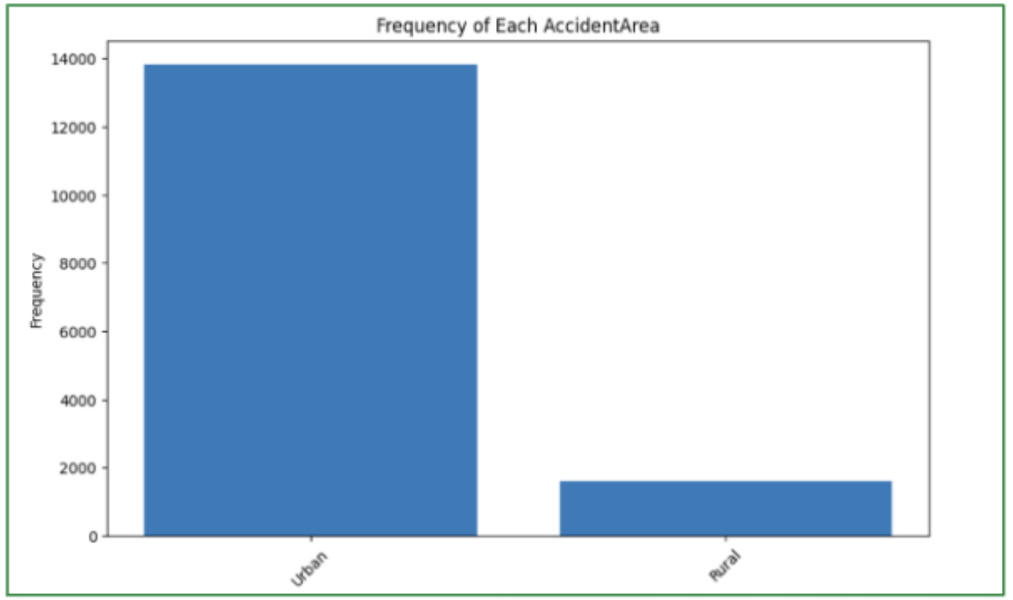
Bivariate Analysis

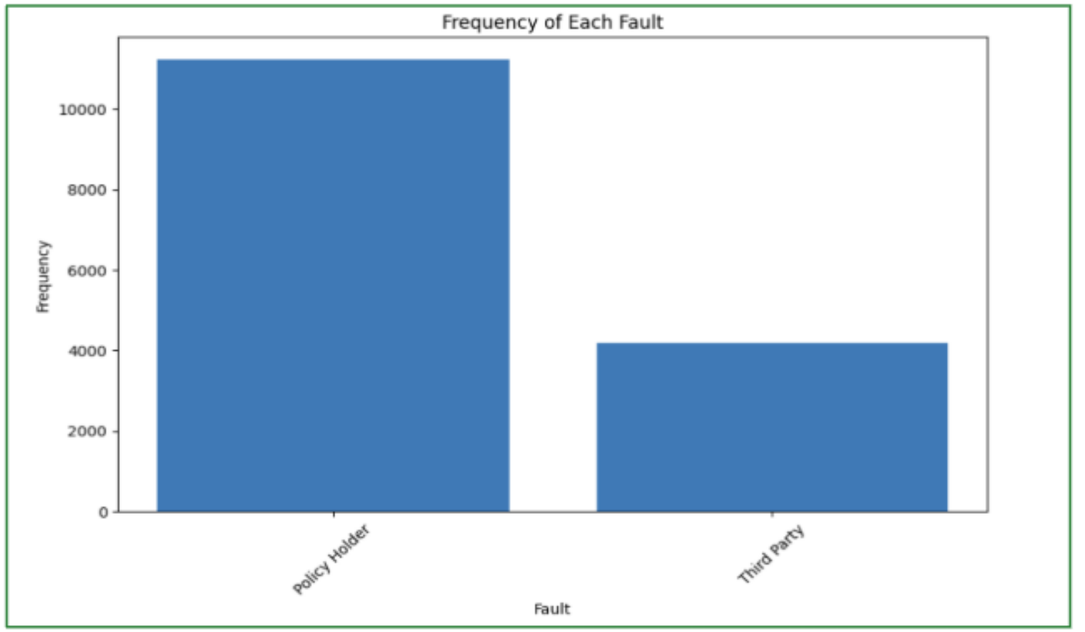
- More urban claims for both genders, and more male policyholders for both rural/urban
- More fraudulent claims among men and married couples
- Fraudulent claims not reported to police

Key Visualizations: Univariate Analysis

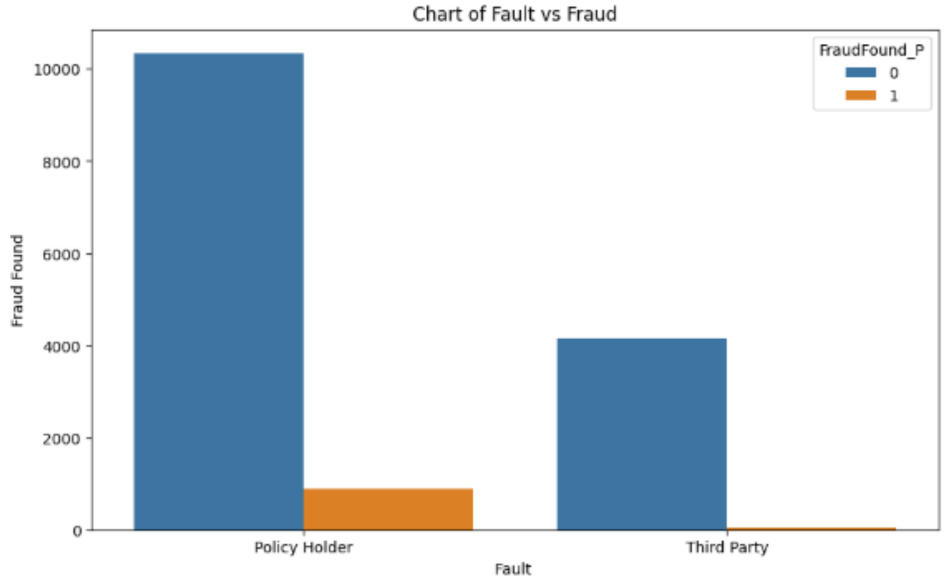


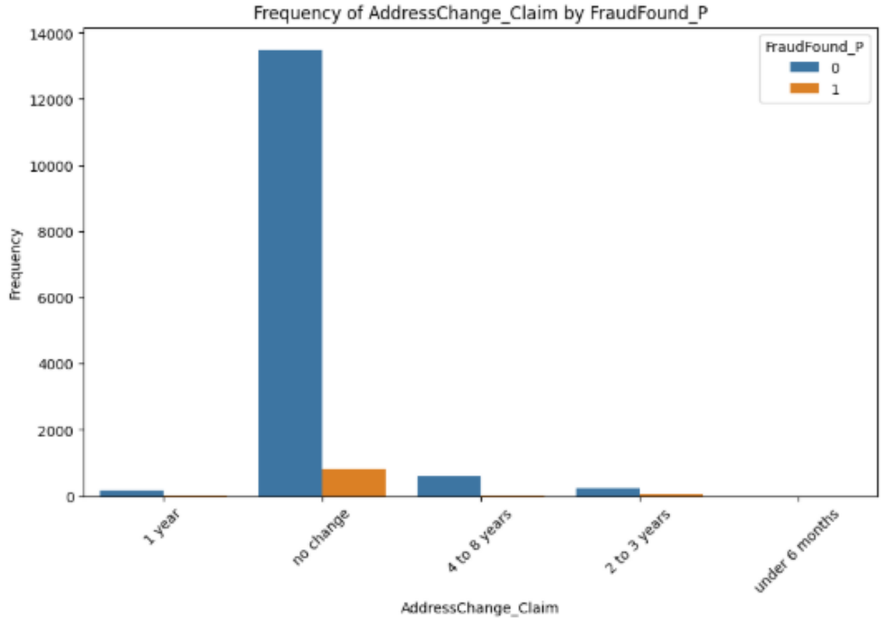


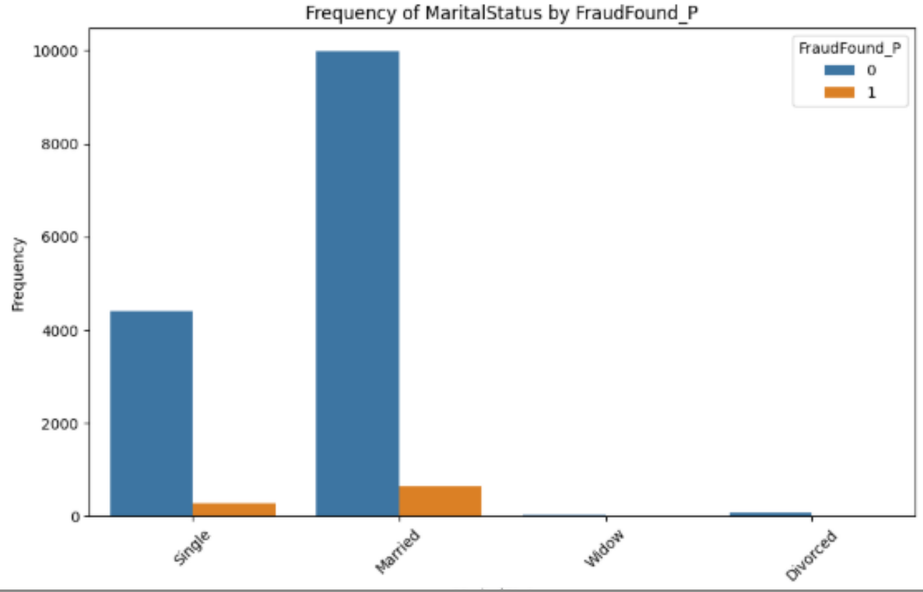


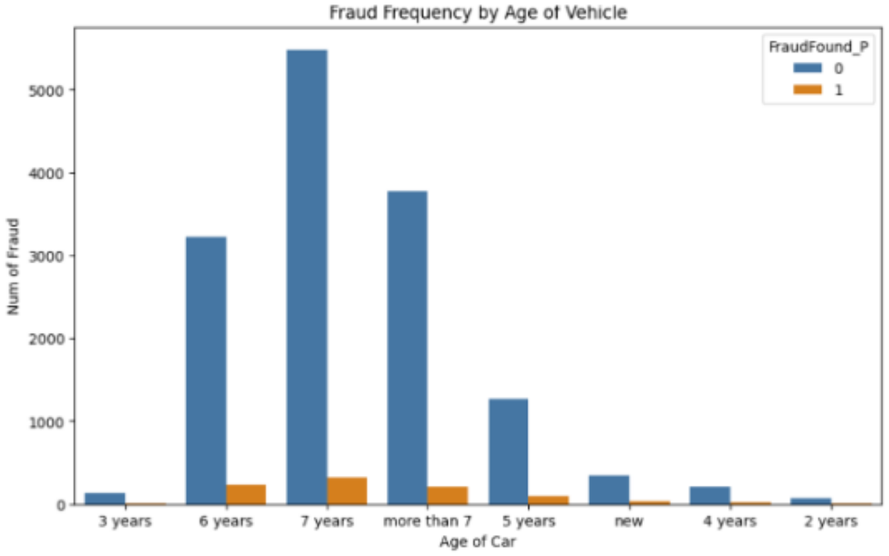


Key Visualizations: Bi-Variate Analysis









Models

Models Chosen:

- Logistic Regression: baseline model to predict binary outcome, interpretable
- KNN: captures nonlinear patterns by assigning class based on nearest data, but sensitive to noise
- SVM: for complex boundaries that best separates classes, kernel flexibility
- Random Forest: bagging with multiple decision trees, handles interactions well, explains feature importance

Evaluation Metrics

- Numerical: Accuracy, Recall, F1 Score
- Visual: Area Under ROC Curve (AUC)

Performance Improvement

- SMOTE method to address class imbalance
- Feature selection for Logistic Regression - reduced features based on coefficients
- Hyperparameter tuning for Random Forest
- Ensemble methods to combine model strengths

Various Models used to perform Analysis:

Logistic Regression: Balanced performance, with a high recall for fraud detection. Useful for a simple model with interpretable results.

KNN: High accuracy but poor recall on the test set for fraud, indicating overfitting to training data.

SVM: High recall but low precision, suggesting that it identifies fraud cases well but with many false positives.

Random Forest: Overfitting on training data, poor recall for fraud, and hyperparameter tuning did not improve the model.

Ensemble Methods: Combining models did not significantly improve recall for fraud detection, suggesting that ensemble methods might not solve the imbalance in the dataset.

AUC: Provides a good comparison of model performance; however, it’s essential to focus on recall for fraud detection.

Logistic Regression

- The Logistic Regression model performs similarly on the testing set (0.64) and training set (0.77) in terms of accuracy, meaning the model does not overfit and generalizes well.
- On both the training and testing data, the model has higher scores for precision on class 0 and recall on class 1, meaning the model has less false negatives of fraud and can accurately identify cases that are not fraud.
- The high recall score on both the training set (0.92) and the testing set (0.90) means the model performs well on identifying fraud.

Redo Logistic Regression after Features Reduced

→ We looked at exponentiated coefficients (odds) for each feature (**Threshold = 0.05**)

→ Dropped the features having coefficients between **1- threshold** and **1+threshold**.

→ Refit the Model with reduced features.

→ The accuracy for this testing data (0.64) is similar to the model's accuracy before dropping any features.

→ The precision for class 0 (0.99) and recall for class 1 (0.90) is also similar to the model's performance before dropping any features.

Conf. Matrix for Training Data

	precision	recall	f1-score	support
0	0.89	0.62	0.73	11598
1	0.71	0.92	0.80	11598
accuracy			0.77	23196
macro avg	0.80	0.77	0.77	23196
weighted avg	0.80	0.77	0.77	23196

Conf. Matrix for Testing Data

	precision	recall	f1-score	support
0	0.99	0.62	0.76	2899
1	0.13	0.90	0.23	185
accuracy			0.64	3084
macro avg	0.56	0.76	0.50	3084
weighted avg	0.94	0.64	0.73	3084

Conf. Matrix for Testing Data with Reduced Features

	precision	recall	f1-score	support
0	0.99	0.62	0.76	2899
1	0.13	0.90	0.23	185
accuracy			0.64	3084
macro avg	0.56	0.76	0.49	3084
weighted avg	0.94	0.64	0.73	3084

Ensemble Methods

- For the ensemble method of **averaging the KNN model, Support Vector Model, and Random Forest model**, the accuracy is **0.86** which suggests the new classifier is relatively accurate on the test data.

- However, the recall score on the class 1 is **0.36**, which suggests somewhat poor performance on predicting cases of fraud.

- F1 score on the ensemble methods for class 1 **(0.24)** indicates a poor balance between precision and recall.

Confusion Matrix for Test Data

	precision	recall	f1-score	support
0	0.96	0.90	0.92	2899
1	0.18	0.36	0.24	185
accuracy			0.86	3084
macro avg	0.57	0.63	0.58	3084
weighted avg	0.91	0.86	0.88	3084

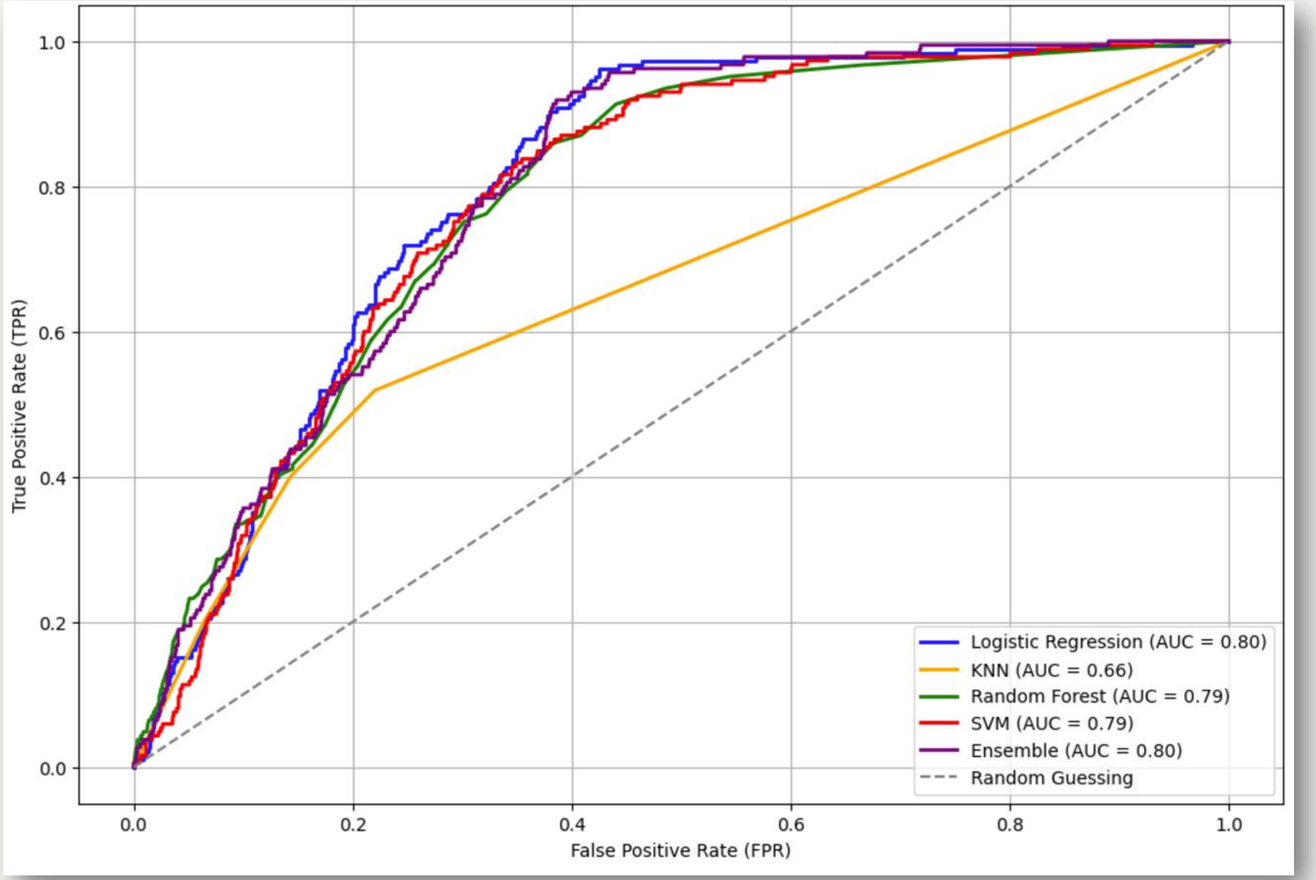
ROC Curve & AUC Scores

- Here we compare five models i.e. LR, KNN, Random Forest, SVM, Ensemble based on their ROC Curve and AUC Curve.

- In the context of vehicle insurance fraud, a **false negative** (or an instance of failing to detect a case of actual fraud) is **more serious** than a false positive (an instance of incorrectly labeling a case as fraud when it is not).

- A false positive may trigger further investigation and potential costs related to such efforts. However, a false negative could greatly cost a policyholder if they are a victim to fraud, or greatly cost the company for paying out fraudulent cases.

- For any model in a business context, in addition to accuracy, we suggest prioritizing **recall** scores to minimize false negatives. Evidently, although the **Ensemble methods** model seems to be the most accurate and has a similar **Area Under the ROC Curve** to logistic regression, the logistic regression has better performance on recall and predicting class 1.



Conclusion

Key Metrics Across Models (Test Set):

Model	Precision (Fraud)	Recall (Fraud)	F1-Score (Fraud)	Accuracy
Logistic Regression	13%	90%	23%	64%
k-NN (k=3)	15%	40%	22%	83%
SVM	14%	71%	24%	73%
Random Forest	22%	13%	16%	92%
Ensemble	18%	36%	24%	86%

Logistic Regression excels in identifying non-fraudulent claims (precision for class 0 is ~99%), making it highly reliable for ruling out legitimate claims.

→ **Business goal:** reducing the investigation workload for non-fraudulent cases.

Why LR?

- **Logistic Regression** achieves the highest **recall**, meaning most fraudulent claims are flagged. Recall is **critical** in fraud detection, as missing fraudulent claims can lead to significant financial losses.

- **For precision (Non-Fraud Cases),** the model demonstrates unmatched precision (~99%) for legitimate claims, allowing the company to confidently automate their clearance.

Prioritize flagged claims using a tiered investigation system (e.g., further review only high-confidence fraud cases).

Business Recommendations

1. The company can implement the model by integrating with existing systems and deploy the model in the company's fraud detection pipeline → **Prioritize flagged claims for manual review by the fraud investigation team.**

2. The business can **allocate investigation resources** to **claims flagged** as high-risk (e.g., above the 0.65 threshold). We can **randomly audit claims** in the 0.2–0.65 range to detect emerging fraud patterns that may not yet be captured fully by the model.

   Approve claims with low fraud risk (e.g., below 0.2) automatically, reducing processing times and improving customer satisfaction.

3. Use the feature importance analysis to identify new variables or refine existing ones. For example, if claim amount deviation and policyholder history are key drivers, these features can be further enhanced with **derived metrics (e.g., loss ratio).**

4. **As the company grows,** the model can scale to handle increased claim volumes without compromising accuracy or speed.

Thank you!