Fraudulent Claim Detection

Submitted by:

Keshav Gupta

Deepika Hegde

Contents

- Problem Statement
- Data Analysis with respect to business objective
- Data Cleaning
- Univariate Analysis
- Bivariate Analysis
- Model Selection

  Logistic Regression

  Random Forest
- Conclusion

Problem Statement

- Global Insure wants to build a model to classify insurance claims as either fraudulent or legitimate based on historical claim details and customer profiles. By using features like claim amount, customer profiles and claim types, the company aims to predict which claims are likely to be fraudulent before they are approved.

- Goal
  - Global Insure aims to enhance its ability to detect fraudulent insurance claims by leveraging historical claim data.
  - The company seeks to identify patterns and key indicators that differentiate fraudulent claims from genuine ones.
  - By developing a predictive model, it intends to assess the likelihood of fraud in incoming claims, enabling proactive fraud detection and reducing financial losses

Data Analysis With Respect To Business Objective

Fraudulent claim detection provided us with data which contained information having 1000 rows and 40 columns related to customer past credit history.

Data Analysis With Respect To Business Objective

Key Analysis

- **Risk Assessment:** Access the risk of fraud based on claim characteristics and claim type.
- **Trend analysis for future prediction:** Identify trends in claim over time(example would be which type of claim are more likely to default i.e. vehicle/property/injury).

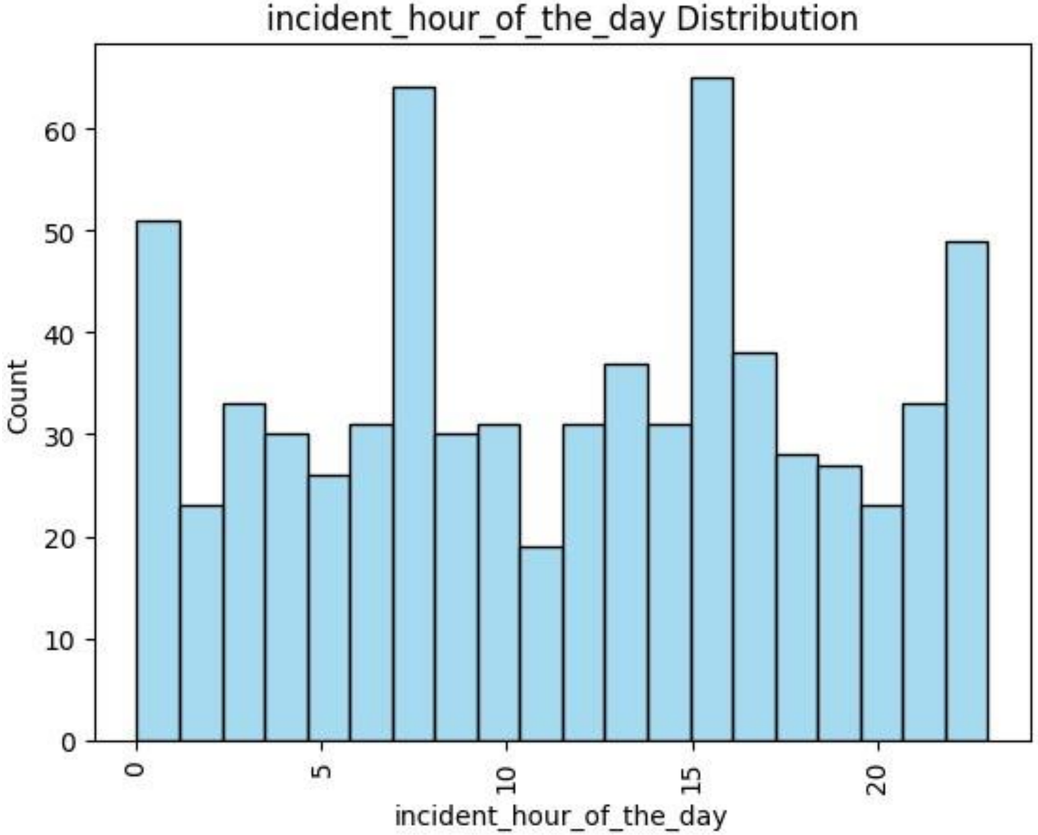
Data Cleaning

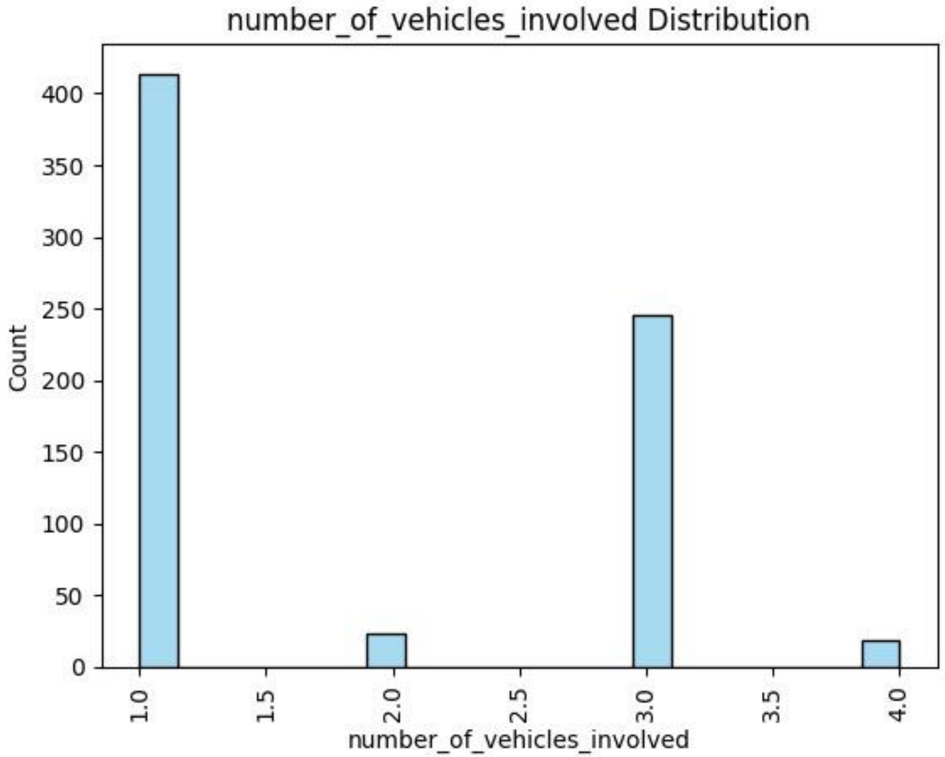
- Replaced the character with missing values for columns

  - ‘property_damage’,‘police_report_available’,‘collision_type’,’authorities.
- Dropped the column which are not adding any value for analysis -C_39,
- Dropped the columns with less predictive power

  -’policy_number’,’insured_hobbies’,’auto_make’,’auto_year’,’incident_location’
- Changed the column format (changed to date-time-format)-’policy_bind_date’,’incident_date’

Univariate Analysis

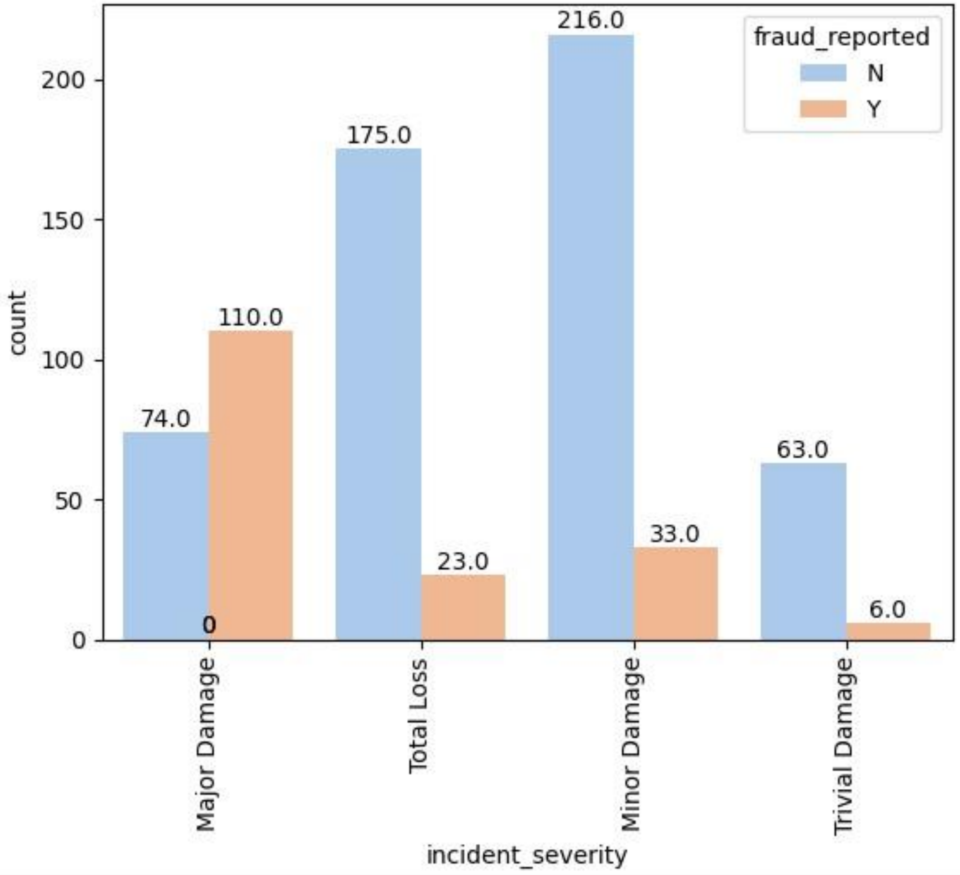


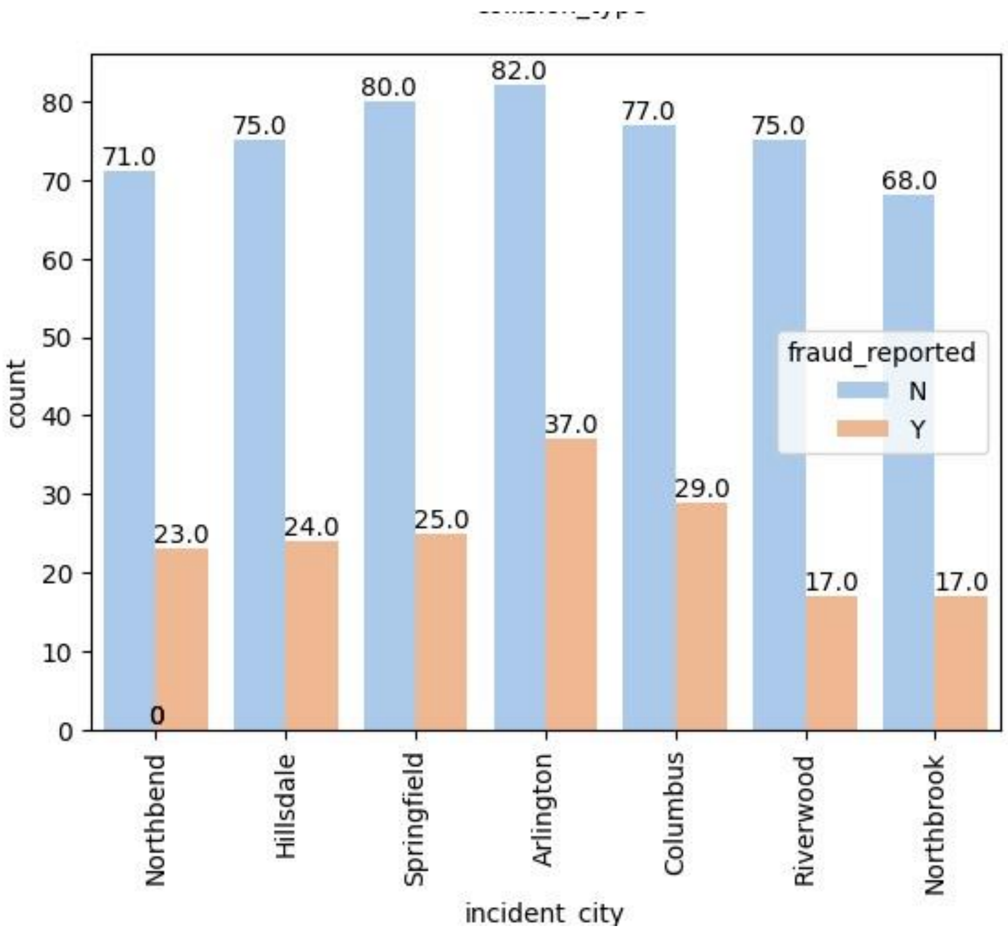


Bivariate Analysis

- Bivariate analysis is an essential part of data analysis which help in identifying the relationship between columns and in decision making process. Type of analysis depend on nature of variables and goal of analysis.
- Analysis was performed between numerical features, Categorical Variables vs target variable.
- Target Variable- 'Fraud_detected'
- Numeric features - 'Age', 'Bodily_injuries', 'Captial_gain', 'Capital_loss', 'Incidnet_date', 'Incident_hour_of_the_day', 'Injury_claim', 'Insured_zip', 'Month_as_customer', 'Number_of_vehicles_involved', 'Policy_annual_premium', 'Property_claim', 'Total_claim_amount'
- Categorical features-’insured_occupation’, ‘insured_relationship’.’incident_type’.’incident_severity’.’authorities_contacted’,’incident_state’.

Bivariate Analysis





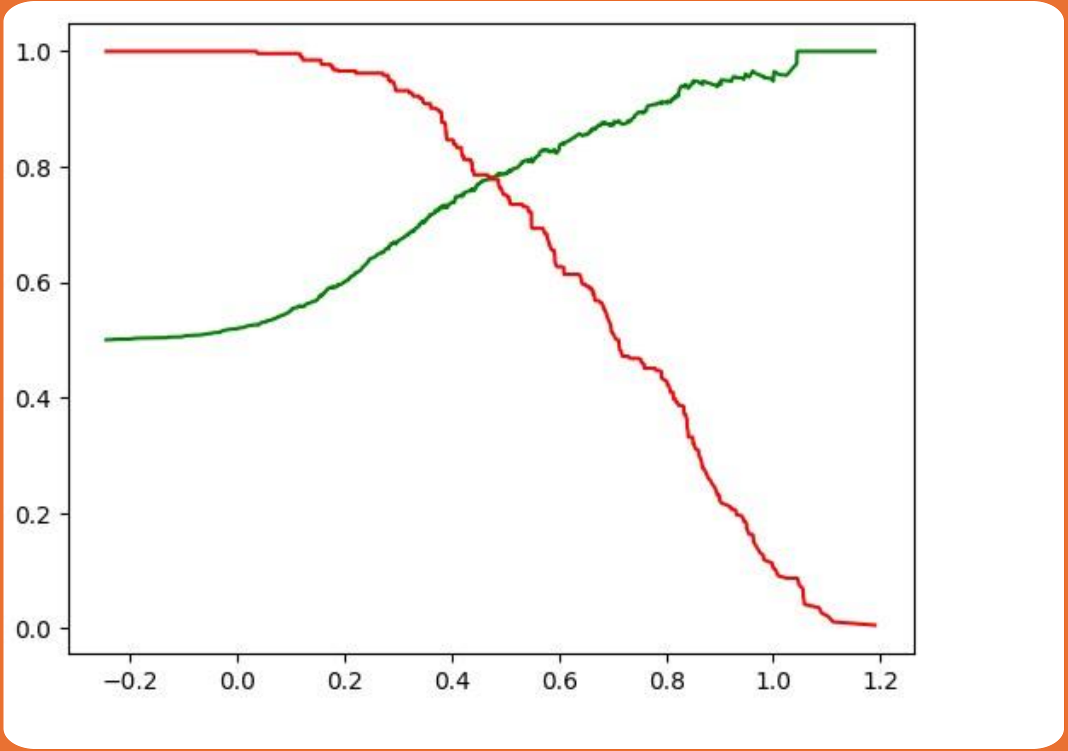
Model Selection

Logistic Regression

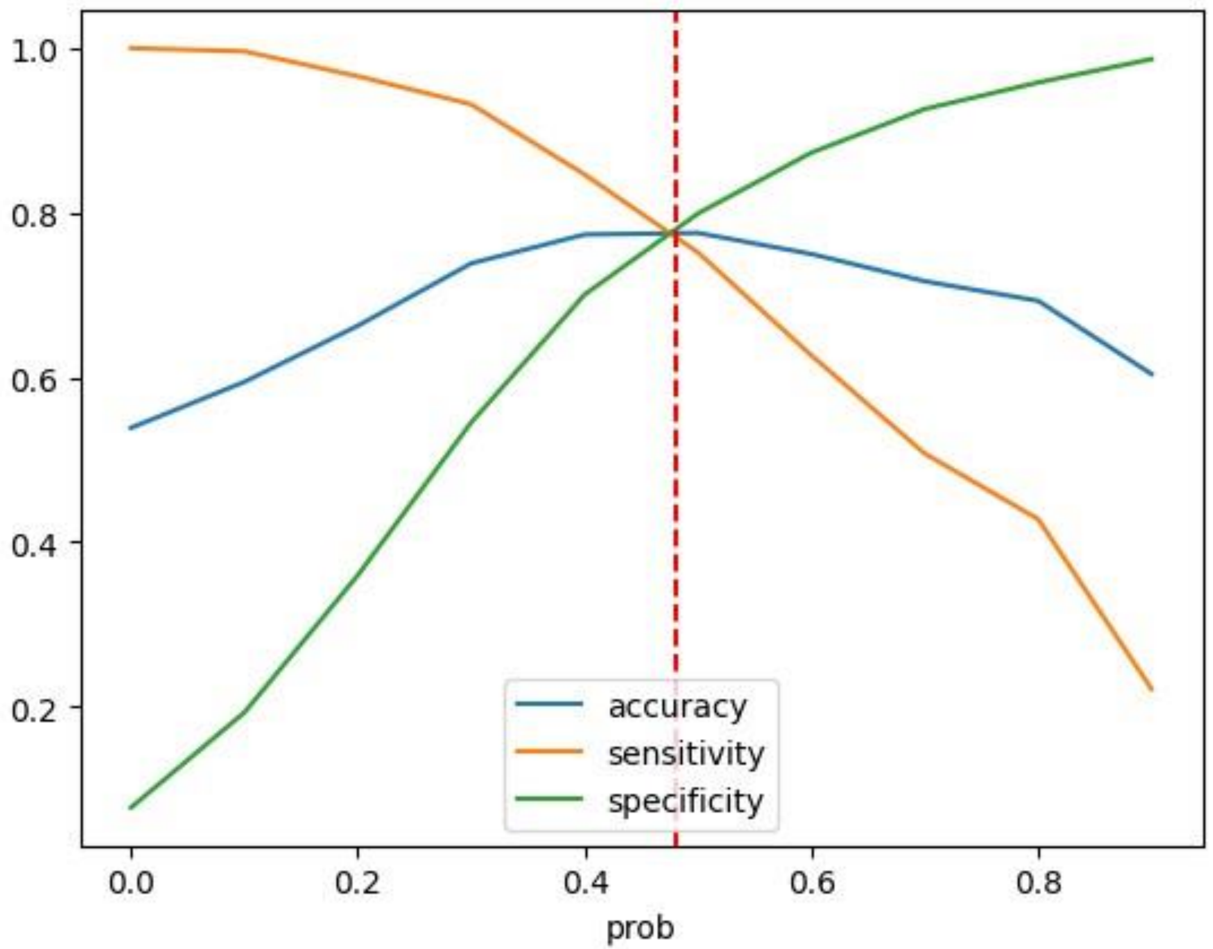
Random Forest

Logistic Regression

Precision-Recall plot



Sensitivity Vs Specificity Vs Accuracy Plot



Model Performance

Metric	Training Data	Validation Data
Precision	83%	57%
Sensitivity	63%	68%
Specificity	87%	83%
Recall	63%	68%
F1 Score	71%	62%

Random Forest

Model Performance

	82%	61%
	77%	44%
	83%	91%
	77%	44%
	80%	51%

Tuned Hyperparameters Random forest

- max_depth - 6
- max_leaf_node - 7
- min_samples_split – 4
- n_estimators -60

Conclusion

- Both the models are overfitting but Logistic regression is performing better compared to Random forest.
- In context of business implications models is likely to catch most fraudulent claims but high false positives would make process inefficient.

Recommendations

- Further analysis should be performed to refine the model and improve its performance.
- The model can be deployed to predict the likelihood of fraud for incoming claims.