Fraudulent Insurance Claim Detection: Final Report

1. Problem Statement

Insurance fraud is a one of the major issue for any industry, resulting in significant financial losses year on year. The objective of this project was to develop a predictive model which is capable of identifying potentially fraudulent insurance claims using the historical data. One of the important challenges is to detect meaningful patterns that distinguish fraudulent claims from genuine ones, also handling class imbalance, and ensuring the model's ability to generalize well to any new data.

2. Methodology

Data Preprocessing

A basic overview of the structure of the dataset is performed before proceeding with any claim analysis. This includes analyzing the data in the claim files to understand their basic structure. This step also includes performing data quality checks like spelling errors, using blanks and null values to denote 'no data' etc. Data preprocesses include missing value treatment, dropping ID and Date columns, and changing non-numeric type columns to chars. The above steps ensured that the dataset was clean and ready for further analysis and modeling.

Train and Validation Split

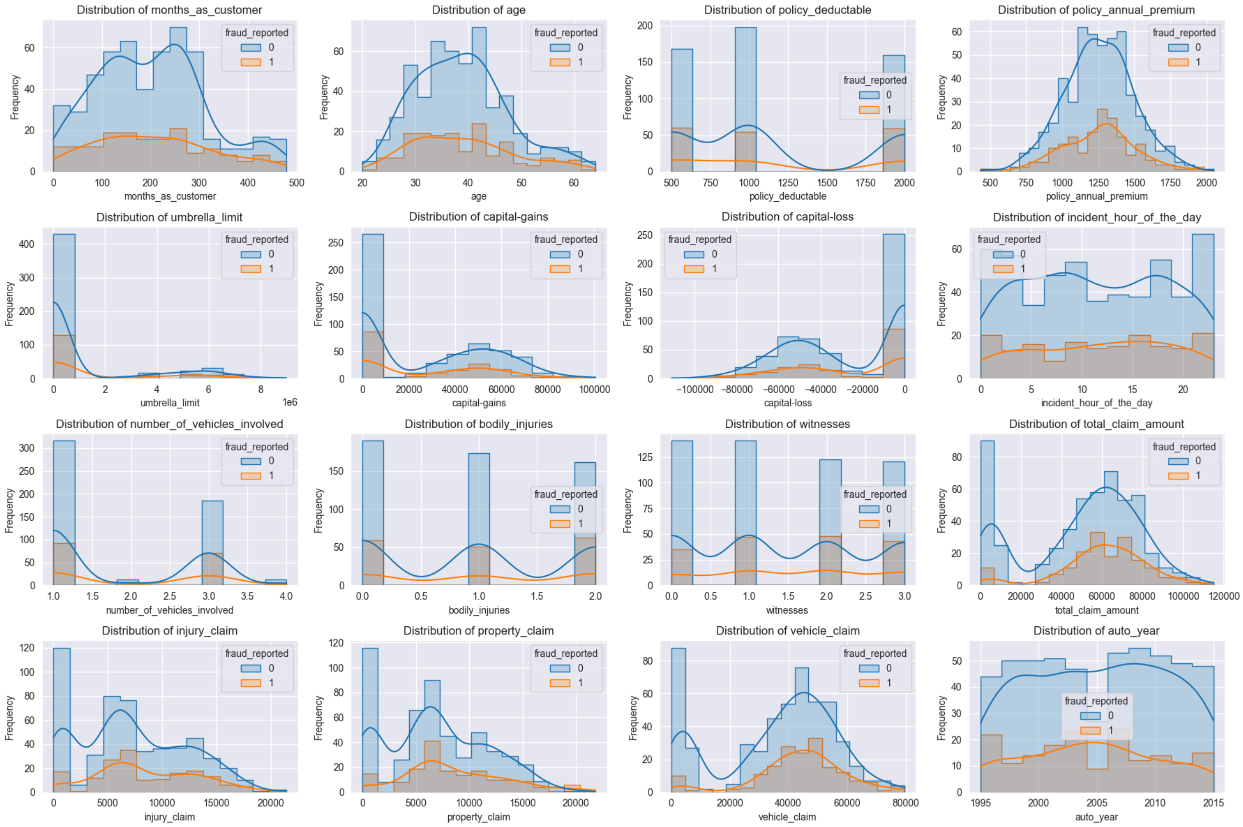
70% of the dataset was allocated as a training set. The remaining 30% was held as a validation set. Stratified sampling was employed to maintain the same ratio of fraudulent cases to non-fraudulent cases in both the training and validation sets. This controlled evaluation and approach prevented data leakage during model evaluation.

Exploratory Data Analysis (EDA)

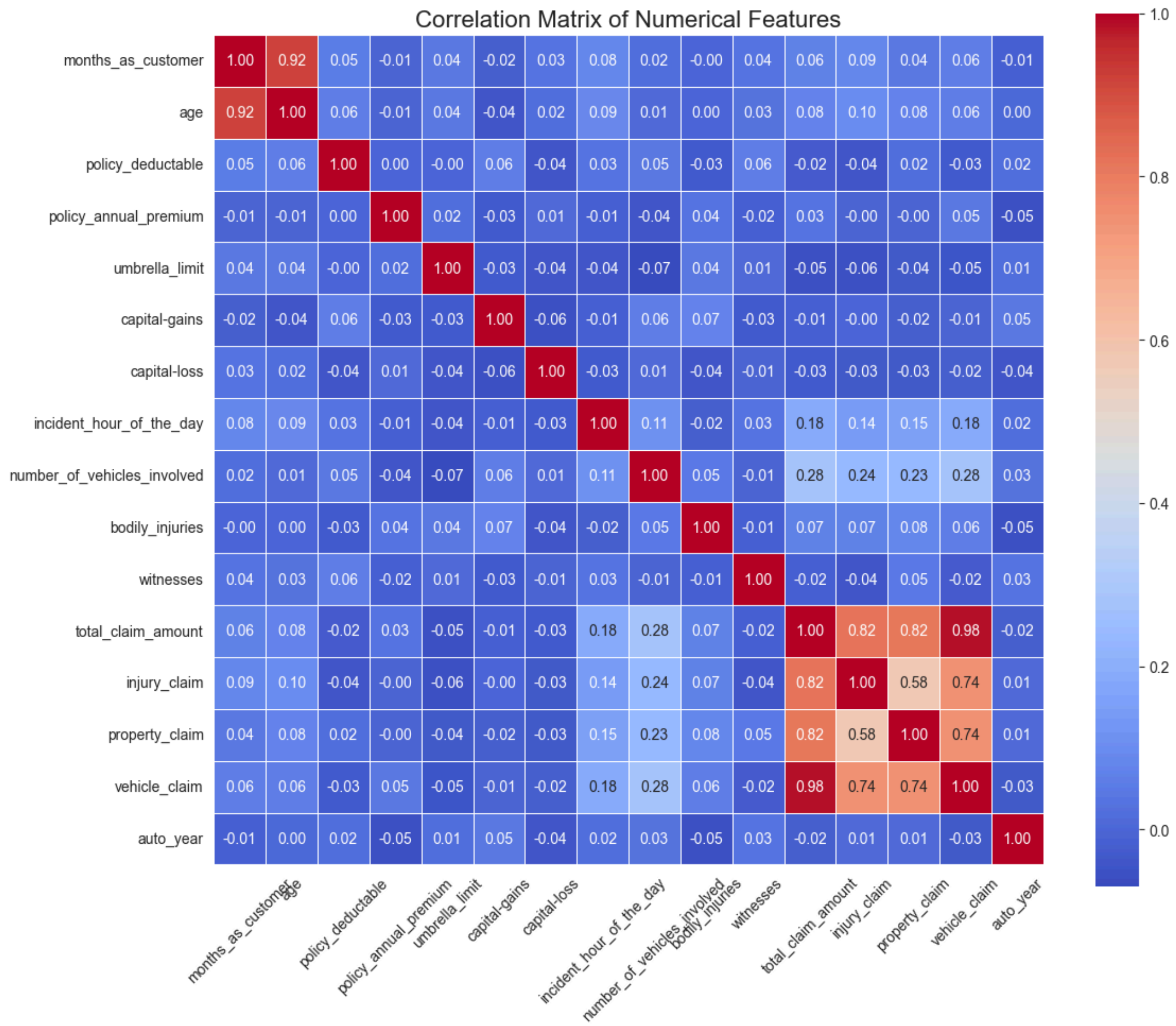
Univariate and bivariate analyses were performed to understand the distribution of features and their relationships with the target variable. Key insights included:

- To gain insights about target variables, univariate and bivariate analyses were executed to observe how features relate to the target variable. Crucial insights included:
- In total there seems to be sub groups within the values of total claim amount, injury claim and property claim variables due to submodal distribution for each of them.

- The normal distribution of policy premium values in relation to policy premium amounts suggests that there is balance around the average and hence policy premiums are normal.
- The propotion of people of age 20 to 70 has a similar number of people of that age thus age proportion is uniformly distributed.
- Frequency of incidents predominantly involved either single or triple vehicle collision.
- Concentration of values for umbrella limit tend to remain at lower level hence showing focus towards it's lower ranges value.



A correlation heatmap was created, highlighting strong correlations between variables such as `total_claim_amount`, `injury_claim`, and `property_claim`, as well as between `age` and `months_as_customer`.



Feature Engineering

Feature engineering steps included:

- Transforming and encoding categorical features to make them applicable and usable for algorithms through the creation of dummy variables.
- Reshaping distribution of data and maintaining consistency with StandardScaler for numerical feature.
- Overcoming class imbalance issues using SMOTE (Synthetic Minority Oversampling Technique) to upsample the minority fraud class in the training set.
- Developing a new feature, claimseverityratio, to quantify the relationship between total claim amount and damage claims.
- Conducting feature selection through Recursive Feature Elimination Cross Validation (RFECV) and Random Forest feature importance determined analysis.

3. Modeling Techniques Used

Logistic Regression:

- Used logistic regression in RFECV to extract informative features.
- Constructed a logistic regression model, then analyzed its performance.
- Evaluated the presence of multicollinearity by calculating the VIF values.
- Refined the probability cut-off based on sensitivity and specificity from the ROC curve analysis.
- The model was evaluated with regard to Accuracy, Precision, Recall, F1-score, Specificity, and ROC–AUC.

Random Forest Classifier:

- Trained the Random Forest model on the selected features.
- Undertook GridSearchCV for hyperparameter tuning of nestimators, maxdepth, and minsamplesleaf.
- Fine-tuned the model after establishing the best parameters.
- Pulled feature importances and retrained the model on a constrained set of features to improve model performance.

4. Evaluation Metrics

- The models are assessed with respect to the following evaluative criteria:
- Accuracy: the model's correctness in all aspects.
- Precision: the actual incidence of fraud on all predicted fraudulent activities.
- Recall (Sensitivity): the ability to detect and report case of fraud.
- Specificity: the ability to identify fraudulent claims that are not true.
- F1 Score: Precision and Recall taken together and averaged using a harmonic mean.
- ROC–AUC: the area under the ROC curve, which is indicative of the model's discrimination ability.

5. Key Insights and Outcomes

- Random forest has a better performance than logistic regression in recall and ROC-AUC as well as other evaluation parameters.
- Detection of fraudulent claim usage was enhanced greatly after using SMOTE to rectify the class imbalance.
- As for the sharp increase in Logistic Regression’s sensitivity, it was due to heightened threshold setting which lessened the chance of missing fraudulent claim detections.
- The most important predictors in building the most predictive model were incident_severity, total_claim_amount, policy_annual_premium and claim_severity_ratio.

- With the added hyperparameter tuning of Random Forest, there was an improvement in detecting sensitive information as well as F1 score, as a result overall performance was increased.

6. Final Recommendation

As for the evaluation performed, the deployment of the Random forest model with a probability cutoff of 0.45 is suggested. It outperforms in turning higher while also not surpassing the threshold of false positive outcomes (high precision and specificity towards true positives). Updating the model periodically, adjusting the threshold, and claim data shifts is recommended. Further optimizations could augment by applying higher-grade parameters such as XGboost/LightGBM or advanced feature construction from a deeper domain perspective.