Case Study : Fraudulent Claim Detection

Poorna Ramakrishnan

Prasita Shetty

Fraudulent Claim Detection

Objective :

Detect fraudulent claims using data driven methods to classify claims as legitimate or fraudulent early in the approval process thereby reducing losses for the company

Questions :

This deck helps to answer the below questions:

- **How can we analyze historical claim data to detect patterns that indicate fraudulent claims?**
- **Which features are most predictive of fraudulent behavior?**
- **Can we predict the likelihood of fraud for an incoming claim, based on past data?**
- **What insights can be drawn from the model that can help in improving the fraud detection process?**

Fraudulent Claim Detection

How can we analyse historical claim data to detect patterns that indicate fraudulent claims? (1/2)

Approach :

To classify claims to be legitimate or fraudulent, a data driven approach was taken:

1. Claims data was analyzed to detect any pattens in it
2. Exploratory data analysis was performed on the data on uncover patterns which includes looking at the variables independently and multiple variables together
3. Two techniques were used to build a model for detecting fraudulent claims

- Logistic Regression
- Random Forest

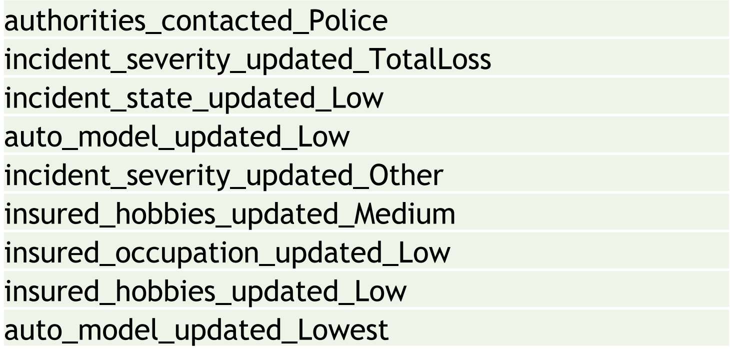
4. The coefficient of the variables in Logistic regression and feature importance from random forest will give us a sense of the important features that influence fraudulent claims
5. Best technique was chosen basis the performance on the test/validation data

Fraudulent Claim Detection

How can we analyse historical claim data to detect patterns that indicate fraudulent claims? (2/2)

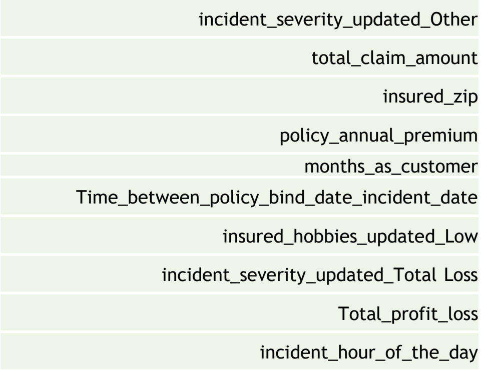
- A data driven approach along with human oversight should be used to check the claims before they are classified as fraudulent or legitimate
- Variables that came out to be important should be taken into consideration when claims of the same profile are to be analysed

Important features from Logistic Regression model (using RFECV)



authorities_contacted_Police
incident_severity_updated_TotalLoss
incident_state_updated_Low
auto_model_updated_Low
incident_severity_updated_Other
insured_hobbies_updated_Medium
insured_occupation_updated_Low
insured_hobbies_updated_Low
auto_model_updated_Lowest

Important features from Random Forest (features whose importance was >.04 were chosen)



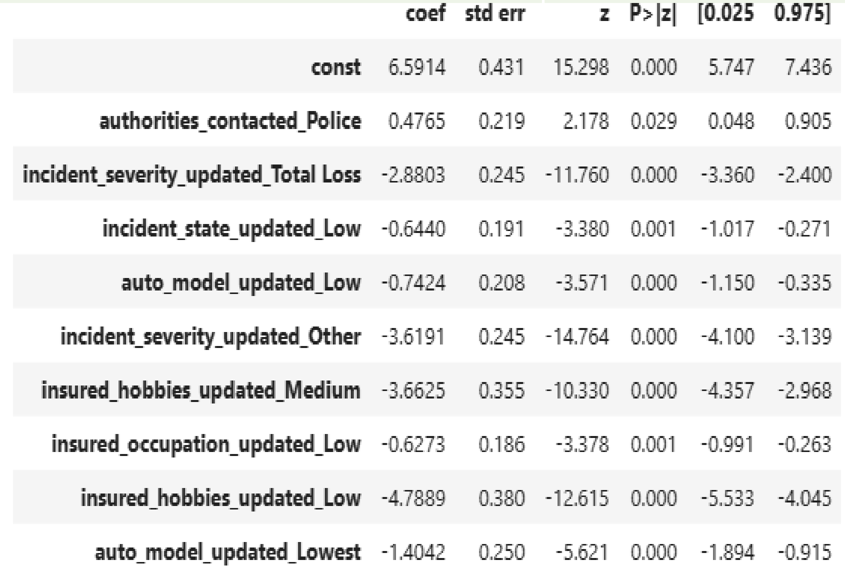
incident_severity_updated_Other
total_claim_amount
insured_zip
policy_annual_premium
months_as_customer
Time_between_policy_bind_date_incident_date
insured_hobbies_updated_Low
incident_severity_updated_Total Loss
Total_profit_loss
incident_hour_of_the_day

- More fraudulent data (if available) can be used for the analysis to make it more robust and understand further patterns

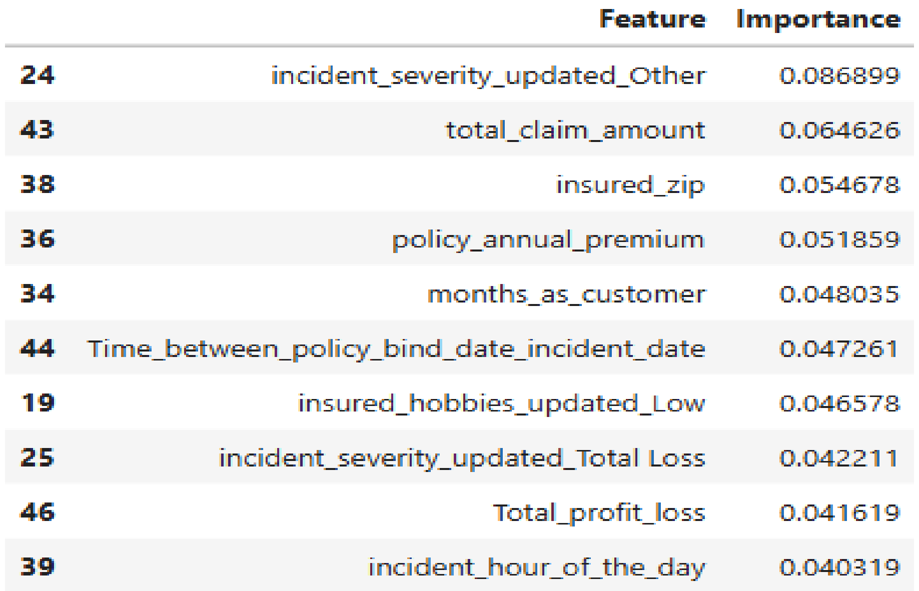
Fraudulent Claim Detection

Features predictive of fraudulent behavior and their interpretation(1/2)

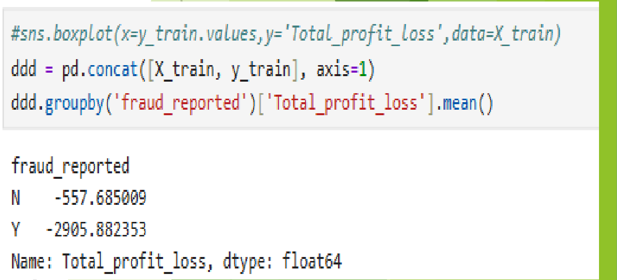
Interpretation	
Authorities_contacted_police	Claims are found to be less fraudulent when police has been contacted
incident_severity_updated_TotalLoss	If the incident severity is 'TotalLoss' , there is less likelihood of that case being fraudulent
incident_state_updated_Low	Claims coming certain states (NY, PA, VA, WA) are likely to be less fraudulent
auto_model_updated_Low	Claims coming from certain auto models (tahoe, wrangler,x6,x5) are likely to be less fraudulent
incident_severity_updated_Other	If the incident severity is 'minor' or 'trivial' damage, then it is less likely to fraudulent
insured_hobbies_updated_Medium	Claims coming from all hobbies are likely to be less fraudulent other than chess and crossfit
insured_occupation_updated_Low	Claims from all occupations except -'exec-managerial','farming-fishing','tech-support','transport-moving','sales','craft-repair' cause less likelihood of frauds
insured_hobbies_updated_Low	Claims coming from all hobbies are likely to be less fraudulent other than chess and crossfit
auto_model_updated_Lowest	Claims coming from certain auto models (legacy, camry, neon, rsx, ultima) are likely to be less fraudulent
Total_profit_loss	Claims are fraudulent when the person has more overall losses from stocks, bonds etc
Total claim amount	Claims (esp vehicle claims) that are either took low in amount or extremely high seem to be more fraudulent than others



	coef	std err	z	P>|z|	[0.025	0.975]
const	6.5914	0.431	15.298	0.000	5.747	7.436
authorities_contacted_Police	0.4765	0.219	2.178	0.029	0.048	0.905
incident_severity_updated_Total Loss	-2.8803	0.245	-11.760	0.000	-3.360	-2.400
incident_state_updated_Low	-0.6440	0.191	-3.380	0.001	-1.017	-0.271
auto_model_updated_Low	-0.7424	0.208	-3.571	0.000	-1.150	-0.335
incident_severity_updated_Other	-3.6191	0.245	-14.764	0.000	-4.100	-3.139
insured_hobbies_updated_Medium	-3.6625	0.355	-10.330	0.000	-4.357	-2.968
insured_occupation_updated_Low	-0.6273	0.186	-3.378	0.001	-0.991	-0.263
insured_hobbies_updated_Low	-4.7889	0.380	-12.615	0.000	-5.533	-4.045
auto_model_updated_Lowest	-1.4042	0.250	-5.621	0.000	-1.894	-0.915

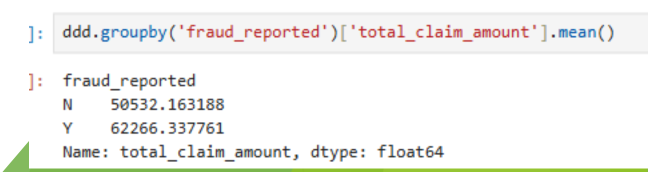


	Feature	Importance
24	incident_severity_updated_Other	0.086899
43	total_claim_amount	0.064626
38	insured_zip	0.054678
36	policy_annual_premium	0.051859
34	months_as_customer	0.048035
44	Time_between_policy_bind_date_incident_date	0.047261
19	insured_hobbies_updated_Low	0.046578
25	incident_severity_updated_Total Loss	0.042211
46	Total_profit_loss	0.041619
39	incident_hour_of_the_day	0.040319



```
#sns.boxplot(x=y_train.values,y='Total_profit_loss',data=X_train)
ddd = pd.concat([X_train, y_train], axis=1)
ddd.groupby('fraud_reported')['Total_profit_loss'].mean()
```

```
fraud_reported
N    -557.685009
Y   -2905.882353
Name: Total_profit_loss, dtype: float64
```



```
]: ddd.groupby('fraud_reported')['total_claim_amount'].mean()

]: fraud_reported
   N    50532.163188
   Y    62266.337761
   Name: total_claim_amount, dtype: float64
```

Fraudulent Claim Detection

Features predictive of fraudulent behavior basis EDA (2/2)

- Ohio has higher % of frauds compared to other 2 states
- Fraud % is high in Ohio(47%)Fraud % is high in Arlington (29%)
- Fraud % is high when policy_csl is 100/300 there are more fraudulent cases in males than females (26%, 23%)
- Frauds are highest in the college group and lowest in Masters group
- Fraud % is high in exec-managerial, farming-fishing and tech-support occupations
- Fraud % is very high where hobbies are chess, cross-fit
- Fraud % is high when the relationship between the insurer is wife
- Fraud % is high in single vehicle collisions
- Fraud % is high in rear collisions
- Fraud % is high when incident severity is 'major damage' (57%)
- Fraud % is higher in cases when ambulance was called
- Fraud % is high in cases where there is no information mentioned for property damage
- Fraud % is high in cases where there is no information mentioned for police_report_available and ‘No’
- Fraud % is high in Meredes,Audi,FordFraud % is high in Silverado, C300 auto models

Fraudulent Claim Detection

Next Steps –

Can we predict the likelihood of fraud for an incoming claim, based on past data?

- **Historical data** is an asset for training fraud detection models.
- Yes, we can use the historical data to predict fraudulent claims , however, the results should be used in conjunction with human verification and validation/ expert human review
- **Continuous monitoring and retraining** of the model are essential to adapt to changing fraud trends and to maintain effectiveness.

Fraudulent Claim Detection

What insights can be drawn from the model that can help in improving the fraud detection process

➢ **Add Contextual Features to Improve Model Robustness**

•Time of claim submission (morning/evening/night)

•Day of week (weekday/weekend)

•Historical claim frequency and pattern

Outcome: Lower false positives and false negatives through behavior-aware models

➢ **Strengthen Documentation and Verification**

•Require reliable proofs (e.g., receipts, timestamped photos)

•Make police verification mandatory for high-severity or high-value claims

Outcome: Stronger evidence base, better fraud deterrence

➢ **Flag and Treat High-Risk Segments Differently**

•Claims from certain geographies or vehicle types

•High-risk hobbies or behavioral indicators

•Unusual claim severity or timing patterns

Outcome: Targeted fraud prevention without penalizing low-risk claims

➢ **Design Tiered Claims Processing Workflows**

•Apply extra checks for flagged segments

•Route low-risk claims through fast-track approvals

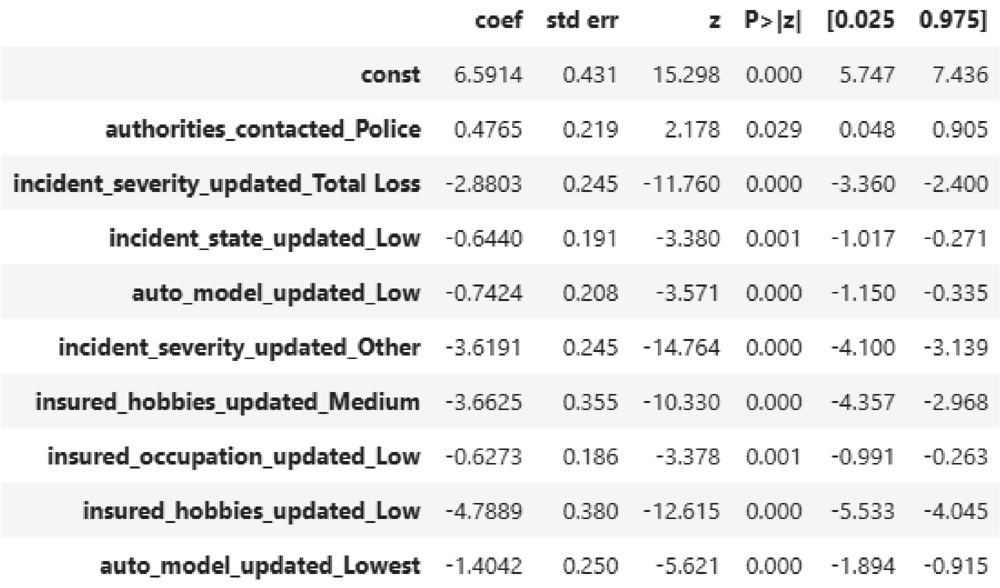
Outcome: Optimized resource use and improved customer experience

APPENDIX

Fraudulent Claim Detection

Recommendations from Logistic regression basis model summary

Interpretation	
incident_severity_updated_TotalLoss	If the incident severity is 'TotalLoss' , there is less likelihood of that case being faudulent
incident_state_updated_Low	Claims coming certain states (NY, PA, VA, WA) are likely to be less fraudulent
auto_model_updated_Low	Claims coming from certain auto models (tahoe, wrangler,x6,x5) are likely to be less fraudulent
incident_severity_updated_Other	If the incident severity is 'minor' or 'trivial' damage, then it is less likely to fraudulent
insured_hobbies_updated_Medium	Claims coming from all hobbies are likely to be less fraudulent other than chess and crossfit
insured_occupation_updated_Low	Claims from all occupations except -'exec-managerial','farming-fishing','tech-support','transport-moving','sales','craft-repair' cause less likelihood of frauds
insured_hobbies_updated_Low	Claims coming from all hobbies are likely to be less fraudulent other than chess and crossfit
auto_model_updated_Lowest	Claims coming from certain auto models (legacy, camry, neon, rsx, ultima) are likely to be less fraudulent



	coef	std err	z	P>|z|	[0.025	0.975]
const	6.5914	0.431	15.298	0.000	5.747	7.436
authorities_contacted_Police	0.4765	0.219	2.178	0.029	0.048	0.905
incident_severity_updated_Total Loss	-2.8803	0.245	-11.760	0.000	-3.360	-2.400
incident_state_updated_Low	-0.6440	0.191	-3.380	0.001	-1.017	-0.271
auto_model_updated_Low	-0.7424	0.208	-3.571	0.000	-1.150	-0.335
incident_severity_updated_Other	-3.6191	0.245	-14.764	0.000	-4.100	-3.139
insured_hobbies_updated_Medium	-3.6625	0.355	-10.330	0.000	-4.357	-2.968
insured_occupation_updated_Low	-0.6273	0.186	-3.378	0.001	-0.991	-0.263
insured_hobbies_updated_Low	-4.7889	0.380	-12.615	0.000	-5.533	-4.045
auto_model_updated_Lowest	-1.4042	0.250	-5.621	0.000	-1.894	-0.915

Fraudulent Claim Detection

Detailed Approach (Analyse claims data to uncover pattens of fraudulent claims)

- Data was segmented into train and evaluation sets
- Numerical , categorical and date variables were analyzed separately
- Derived variables/interaction variables were created where necessary
- Standardization was performed on the numerical data before modelling and its relationship with target variables observed
- Every level of a categorical variable was analyzed with the target variables to see the distribution under each level
- Levels in the categorical variables was combined basis their distribution of fraudulent cases in each case
- All the above transformations were fit only using the train set to avoid data leakage
- The transformed data was fed into 2 different algorithms – Logistic Regression and Random Forest
- The performance of the algorithm was evaluated basis the validation set and the final model was selected

Fraudulent Claim Detection

Methodology

- Data Cleaning was performed
  - Remove redundant columns, impute missing values, change the datatype of the variables as necessary
  - Multicollinearity was identified using heat maps and those variables were dropped
- Data was segmented into train and evaluation sets
  - A 70-30 split was used to segment the data into stratified train and test sets.
  - Class balance was 75-25. Hence, the train data was up sampled to ensure that the 2 target classes are balanced
- Numerical , categorical and date variables were analyzed separately
  - Numerical variables – Standardization technique was used to scale the numerical variables so that no variable is given more importance than the other. This is especially important for Logistic regression
  - Categorical variables- There were 2 type of categorical variables: Variables with less than 3 levels – these were dummy value encoded directly.
  - For the categorical variables , with many levels, the fraud distribution for each level of the variable was observed . Based on this , this, the similar values were combined thus decreasing the cardinality of the variable . Ex: Hobbies, Education , Type, Severity etc. Dummy variable encoding was used on the rolled-up values.

Fraudulent Claim Detection

Methodology

- Derived variables/interaction variables were created where necessary
  - Month and year features were extracted from datetime variables
  - Capital Profile and loss variables were added for a particular candidate to get their overall profile/loss
- Standardization was performed on the numerical data before modelling and its relationship with target variables observed
  - Standard Scalar was performed on the numerical variables to scale the variables to bring them to the same range
- All the above transformations were fit only using the train set to avoid data leakage
- The transformed data was fed into 2 different algorithms – Logistic Regression and Random Forest
- The performance of the algorithm was evaluated basis the validation set and the final model was selected

Fraudulent Claim Detection

Methodology

- The transformed data was fed into 2 different algorithms – Logistic Regression and Random Forest
- All the above transformations were fit only using the train set to avoid data leakage
- Logistic Regression was first tried:
- RFECV (Recursive feature elimination along with cross validation) was used to select the important features. A CV fold of 5 was used .
- P values of the selected variables were observed and VIF was also calculated to ensure there is no multi collinearity on the selected variables
- Predictions were calculated for different threshold values and then the best threshold value was selected based on the various metrics like accuracy, sensitivity and specificity
- ROC curve was plotted to understand AUC
- Accuracy, Sensitivity and specificity was also plotted across various threshold values to confirm the optimal threshold value visually
- Precision recall curve was also plotted to understand these values.
- Random Forest model was then tried:
- A random forest model was calculated with default hyperparameters. This model overfit on the data
- GridSearch CV was then used to get the best hyperparameters
- The selected hyperparameters was then used to construct the final model.
- Top features (whose threshold value >.04) was selected to build the model

Observations:

- Logistic regression model had similar accuracy on train and test and did not overfit
- Random Forest default model overfit here and tuned model performed better on the train set but again overfit on the test set

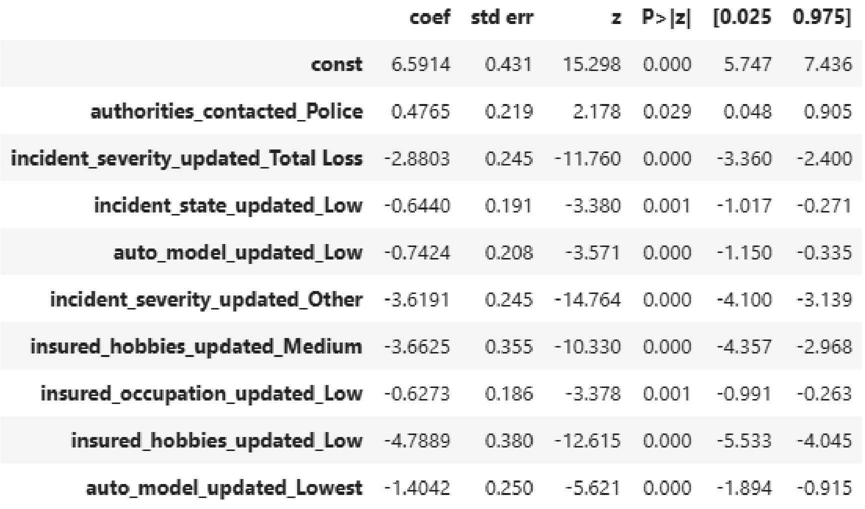
Fraudulent Claim Detection

Reasons for selection of Logistic Regression as the final model

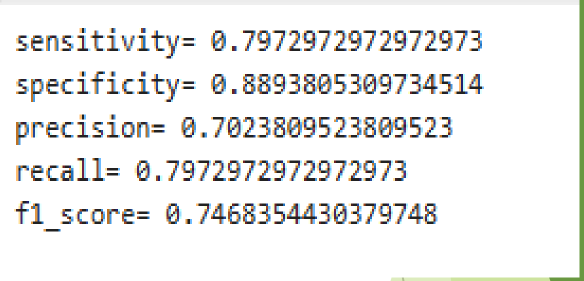
- Logistic regression model was selected as the best model due to its
  - Better performance on the evaluation set
  - Interpretability of the model coefficients and confidence basis p-values

Logistic Regression

accuracy=87%
sensitivity= 79%
specificity= 89%
precision= 70%
recall= 80%
f1_score= 75%



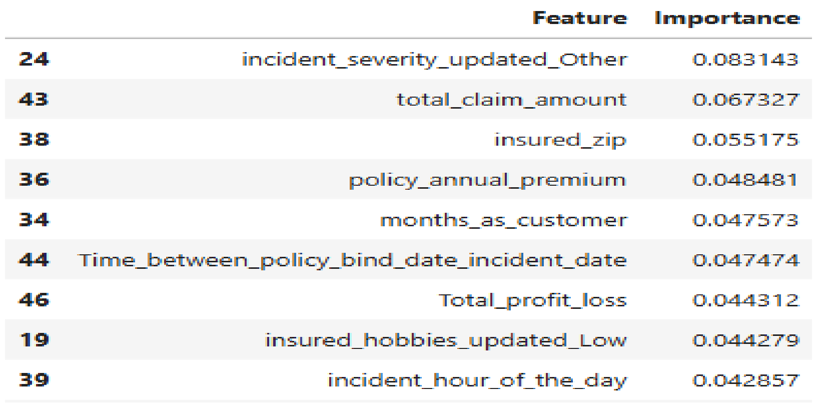
	coef	std err	z	P>|z|	[0.025	0.975]
const	6.5914	0.431	15.298	0.000	5.747	7.436
authorities_contacted_Police	0.4765	0.219	2.178	0.029	0.048	0.905
incident_severity_updated_Total Loss	-2.8803	0.245	-11.760	0.000	-3.360	-2.400
incident_state_updated_Low	-0.6440	0.191	-3.380	0.001	-1.017	-0.271
auto_model_updated_Low	-0.7424	0.208	-3.571	0.000	-1.150	-0.335
incident_severity_updated_Other	-3.6191	0.245	-14.764	0.000	-4.100	-3.139
insured_hobbies_updated_Medium	-3.6625	0.355	-10.330	0.000	-4.357	-2.968
insured_occupation_updated_Low	-0.6273	0.186	-3.378	0.001	-0.991	-0.263
insured_hobbies_updated_Low	-4.7889	0.380	-12.615	0.000	-5.533	-4.045
auto_model_updated_Lowest	-1.4042	0.250	-5.621	0.000	-1.894	-0.915



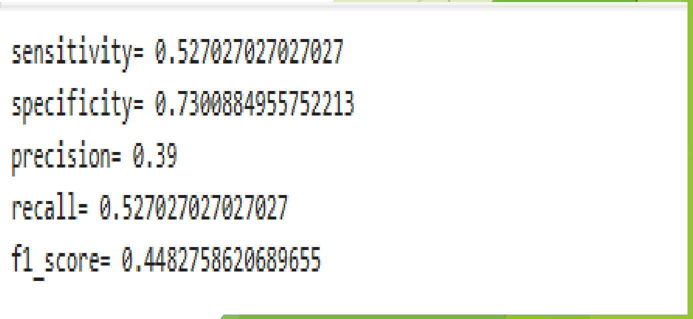
```
sensitivity= 0.7972972972972973
specificity= 0.8893805309734514
precision= 0.7023809523809523
recall= 0.7972972972972973
f1_score= 0.7468354430379748
```

Random Forest

accuracy=68%
sensitivity= 53%
specificity= 73%
precision= 39%
recall= 53%
f1_score= 45%



	Feature	Importance
24	incident_severity_updated_Other	0.083143
43	total_claim_amount	0.067327
38	insured_zip	0.055175
36	policy_annual_premium	0.048481
34	months_as_customer	0.047573
44	Time_between_policy_bind_date_incident_date	0.047474
46	Total_profit_loss	0.044312
19	insured_hobbies_updated_Low	0.044279
39	incident_hour_of_the_day	0.042857



```
sensitivity= 0.527027027027027
specificity= 0.7300884955752213
precision= 0.39
recall= 0.527027027027027
f1_score= 0.4482758620689655
```