**Article on Evaluation Project**

**Project name: Insurance Claim Fraud Detection**

1. **Problem Definition**

Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem.

In this project the dataset provided has the details of the insurance policy along with the customer details. It also has the details of the accident on the basis of which the claims have been made.

In this project a predictive model is created that can predict if an insurance claim is fraudulent or not.

**Dataset Link provided:**

https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/Automobile\_insurance\_fraud.csv

1. **Data Analysis**

**Dataset Overview:**

* The dataset contains 1000 rows and 40 columns.
* Binary classification problem - 'fraud\_reported' (Yes/No).

**Columns Descriptions:**

Columns include the following various features:

1. months\_as\_customer: Number of months of patronage
2. age: the length of time a customer has lived or a thing has existed
3. policy\_number: It is a unique id given to the customer, to track the subscription status and other details of customer
4. policy\_bind\_date:date which document that is given to customer after we accept your proposal for insurance
5. policy\_state: This identifies who is the insured, what risks or property are covered, the policy limits, and the policy period
6. policy\_csl: is basically Combined Single Limit
7. policy\_deductable: the amount of money that a customer is responsible for paying toward an insured loss
8. policy\_annual\_premium: This means the amount of Regular Premium payable by the Policyholder in a Policy Year
9. umbrella\_limit: This means extra insurance that provides protection beyond existing limits and coverages of other policies
10. insured\_zip: It is the zip code where the insurance was made
11. insured\_sex: This refres to either of the two main categories (male and female) into which customer are divided on the basis of their reproductive functions
12. insured\_education\_level: This refers to the Level of education of the customer
13. insured\_occupation: This refers Occupation of the customer
14. insured\_hobbies: This refers to an activity done regularly by customer in his/her leisure time for pleasure.
15. insured\_relationship: This whether customer is: single; or. married; or. in a de facto relationship (that is, living together but not married); or. in a civil partnership
16. capital-gains: This refers to profit accrued due to insurance premium
17. capital-loss: This refers to the losses incurred due to insurance claims
18. incident\_date: This refers to the date which claims where made by customers
19. incident\_type: This refers to the type of claim/vehicle damage made by customer
20. collision\_type: This refers to the area of damage on the vehicle
21. incident\_severity: This refers to the extent/level of damage
22. authorities\_contacted: This refers to the government agencies that were contacted after damage
23. incident\_state: This refers to the state at which the accident happened
24. incident\_city: This refers to the city at which the accident happened
25. 1ncident\_location: This refers to the location at which the accident happened
26. incident\_hour\_of\_the\_day: The period of the day which accident took place
27. number\_of\_vehicles\_involved: This refers to number of vehicles involved the accident
28. property\_damage: This refers to whether property was damaged or not
29. bodily\_injuries: This refers to injuries sustained
30. witnesses: This refers to the number of witnesses involved
31. police\_report\_available: This refers to whether the report on damage was documented or not
32. total\_claim\_amount: This refers to the financial implications involved in claims
33. injury\_claim: This refers to physical injuries sustained
34. property\_claim: This refers to property damages during incident
35. vehicle\_claim: This refers to property damages during incident
36. auto\_make: This refers to the make of the vehicle
37. auto\_model: This refers to the model of the vehicle
38. auto\_year: This refers to the year which the vehicle was manufactured
39. \_c39:
40. fraud\_reported

**The data types in the provided dataset:**

* object – categorical data
* float64 – numerical data
* int64 – numerical data

**Missing and Duplicate Values Analysis:**

* Identified 1000 missing values represented as NaN in ‘\_c39’ column.
* Identified missing values represented as "?" and replaced with NaN in the following columns – ‘collision\_type’, ‘property\_damage’ and ‘police\_report\_available’.
* No duplicate values had been identified.

During the data analysis process the number of unique values had been identified, as well as their count (for reference check the Jupiter notebook file).

1. **Exploratory Data Analysis (EDA) Concluding Remarks**

* Identified 1000 missing values in ‘\_c39’ column – that had been investigated and the ‘\_c39’ had been dropped as it does not contain any values.
* Identified missing values in ‘collision\_type’, ‘property\_damage’ and ‘police\_report\_available’ - Imputed missing values for those features using Simple Imputer immutation.
* Separation of numerical and categorical data had been performed.

Categorical Columns: ['policy\_bind\_date', 'policy\_state', 'policy\_csl', 'insured\_sex', 'insured\_education\_level', 'insured\_occupation', 'insured\_hobbies', 'insured\_relationship', 'incident\_date', 'incident\_type', 'collision\_type', 'incident\_severity', 'authorities\_contacted', 'incident\_state', 'incident\_city', 'incident\_location', 'property\_damage', 'police\_report\_available', 'auto\_make', 'auto\_model', 'fraud\_reported']

Numerical Columns: ['months\_as\_customer', 'age', 'policy\_number', 'policy\_deductable', 'policy\_annual\_premium', 'umbrella\_limit', 'insured\_zip', 'capital-gains', 'capital-loss', 'incident\_hour\_of\_the\_day', 'number\_of\_vehicles\_involved', 'bodily\_injuries', 'witnesses', 'total\_claim\_amount', 'injury\_claim', 'property\_claim', 'vehicle\_claim', 'auto\_year']

* Following the removal of the missing values in the specified column ‘\_c39’ the dataset now has 1000 rows and 39 columns.

During the data visualization we made the following observations:

A group of colorful bars

Description automatically generated with medium confidence

1. **Policy State:**

* The dataset includes three policy states: Illinois (IL), Indiana (IN), and Ohio (OH).
* Ohio (OH) has the highest count - 352, followed by Illinois (IL) – 338 and Indiana (IN) – 310.

1. **Policy csl:**

* The dataset includes three Combined Single Limits – 250/500, 100/300, 500/1000.
* 250/500 has the highest count - 351, followed by 100/300 with count 349 and 500/1000 with count 300.

1. **Insured Sex:**

* The dataset includes two values: Female and Male.
* Female has the highest count – 537 and Male – 463.

1. **Insured Education Level:**

* The dataset includes seven values: JD, High School, MD, Masters, PhD and College.
* JD - 161, High School - 160, Associate - 145, MD - 144, Masters - 143, PhD - 125, College – 122.

1. **Insured Occupation:**

* Various occupations are represented in the dataset.
* ‘machine-op-inspct’ and ‘prof-specialty’ higher counts – respectively 93 and 85 each.
* ‘handlers-cleaners’ and ‘farming-fishing’ have lower counts – respectively 54 and 53 each.

1. **Insured Hobbies:**

* Diverse hobbies are present in the dataset.
* ‘Reading’, 'Exercise' and 'Paintball' have the highest counts – respectively 64 and 57 each.
* 'Basketbal,' and 'Cross-fit' have lower counts – respectively 35 and 34.

1. **Insured Relationship:**

* Various relationships are presented in the dataset, with 'Own-child' and 'Other-relative' having the highest counts – 183 and 177 respectively.
* 'Wife' and 'Unmarried' have the lowest count – 155 and 141 respectively.

1. **Incident Type:**

* The majority of incidents are of the type 'Multi-vehicle Collision' - 419 and 'Single Vehicle Collision' - 403.
* 'Parked Car' incidents have a lower count – 84 and 'Vehicle Theft' incidents - 94.

1. **Collision Type:**

* The dataset includes three collision types: 'Front Collision,' 'Rear Collision,' and 'Side Collision.'
* 'Rear Collision' has the highest count - 470, followed by 'Side Collision' - 276 and 'Front Collision' - 254.

1. **Incident Severity:**

* The dataset includes four categories: ‘Minor Damage’, ‘Total Loss’, ‘Major Damage’ and ‘Trivial Damage’.
* 'Minor Damage' has the highest count - 354, while 'Trivial Damage' has the lowest - 90.

1. **Authorities Contacted:**

* The dataset includes five categories – ‘Police’, ‘Fire, ‘Other’, ‘Ambulance’ and ‘None’.
* 'Police' and 'Fire' have the highest count – respectively 292 and 223, while 'None' lowest - 91.

1. **Incident State:**

* The dataset includes seven states – ‘NY’, ‘SC’, ‘WV’, ‘VA’, ‘NC’, ‘PA’ and ‘OH’.
* ‘NY’ and ‘SC’ have the highest count of incidents – 262 and 148 respectively. ‘PA’ and ‘OH’ have the lowest occurrence of incidents – 30 and 23 respectively.

1. **Incident City:**

* 'Springfield' and 'Arlington' have higher incident counts, each 157 and 152 respectively.

1. **Property Damage:**

* The majority of cases do not involve property damage - 698, while 302 cases involve property damage.

1. **Police Report Available:**

* Around 686 of the cases do not have a police report available, while 314 have a police report.

1. **Auto Make and Model:**

* Various auto makes and models are represented in the dataset, each of them with different count.

1. **Fraud Reported:**

* The dataset is imbalanced regarding fraud reporting. Approximately 753 of cases are reported as 'N' (No fraud), and 247 are reported as 'Y' (Fraud reported).

A group of blue and white graphs

Description automatically generated

1. **Months as Customer:**

* The count of the months as a customer varies.
* The majority of the customers been for 294 months with the company.

1. **Age:**

* The ages are diverse, ranging from 19 to 64.
* The peak is around 43 and 39.

1. **Policy Deductible:**

* The Policy Deductible has three categories – 500, 1000 and 2000.
* The highest count has 1000 with 351 occurrences, followed by 500 with 342 and 2000 with 307

1. **Umbrella Limit:**

* There is wide range of umbrella limits in the dataset.
* A majority of the cases 798 have an umbrella limit of 0.

1. **Capital Gains and Losses:**

* The dataset has wide range of Capital Gain and Losses.
* The most common value for Capital Gain is 0 with 508 occurrences and 0 for Capital Loss with 475 occurrences.

1. **Incident Hour of the Day:**

* The incidents happened different time during the day and night.
* The peak is 17 o`clock, also between 23, 0 and 3 in the morning.

1. **Number of Vehicles Involved:**

* The majority of the incidents had 1 vehicle involved – 581 cases and 3 vehicles – 358 cases.
* The least cases occurred with 4 vehicles – 31 and 2 cars – 30.

1. **Bodily Injuries :**

* The data distribution is quite even

1. **Witnesses:**

* The data distribution is quite even.

1. **Claim Amounts (Injury, Property, Vehicle):**

* The dataset has wide range of the Claim Amounts.

1. **Auto Year:**

* Auto years range from 1995 to 2015.
* Most occurrences are in the years 1995, 1999, and 2005.

Comparing the outcome ‘Fraud\_reported’ with the rest of the columns

A group of graphs and charts

Description automatically generated with medium confidence

**1. Policy State:**

* In all of the states (IL, IN, OH) majority of the reported cases have been non-fraudulent.
* Policy state OH has the highest count of fraud reports - 91.

**2. Insured Sex:**

* Among non-fraudulent cases (N), there are more female policyholders.
* In fraudulent cases (Y), the count is relatively similar.

3. **Insured Education Level:**

* For both cases fraudulent and non-fraudulent the data distribution is quite even.
* The highest number of fraudulent cases reported are with education level JD – 42 cases. The lowest number are Master and College, both 32 cases.
* The highest number of non-fraudulent cases reported are with education level High-school – 124 cases. The lowest number College.

4. **Insured Occupation:**

* In non-fraudulent cases, occupations like adm-clerical, armed-forces, and craft-repair are more common.
* In fraudulent cases, there is a shift towards occupations like exec-managerial, tech-support, and sales.

5. **Insured Hobbies:**

* In non-fraudulent cases, hobbies like base-jumping, camping, and dancing are more prevalent.
* In fraudulent cases, there is a shift towards hobbies like chess, cross-fit, and skydiving.

6. **Insured Relationship:**

* In non-fraudulent cases ‘own child’ has the highest count - 144 cases.
* In fraudulent cases ‘other relative’ has the highest count – 52 cases

7. **Incident Type:**

* Multi-vehicle collisions and single-vehicle collisions are more common in both non-fraudulent and fraudulent cases.
* Parked car incidents are less common but have a higher percentage in non-fraudulent cases.

8. **Collision Type:**

* Rear collisions are the most common in both non-fraudulent and fraudulent cases.

9. **Incident Severity:**

* In non-fraudulent cases minor damage and total loss incidents are more common.
* In fraudulent cases major damage incidents are more common.

10. **Authorities Contacted:**

* Most cases, both fraudulent and non-fraudulent, involve contacting the police.

11. **Incident State and City:**

* The data distribution of incident states and cities varies, with different cities having different count of fraud.

12. **Property Damage:**

* Non-fraudulent cases have higher count where there is no property damage.
* The fraudulent cases have higher percentage where there is a property damage.

13. **Police Report Available:**

* Non-fraudulent cases have higher count where there is no report available.
* The fraudulent cases have higher percentage where there is police report.

14. **Auto Make and Model:**

* The data distribution of auto make and model varies.

A screenshot of a graph

Description automatically generated

1**. Months as Customer:**

* Majority of fraud reported cases occurred in the early months of customer relationships (e.g., first few months).

2. **Age:**

* The majority of reported fraud cases seem to involve younger individuals.

3. **Policy Deductible:**

* Lower policy deductibles – 500 have a higher proportion of fraud reported cases compared to higher deductibles.

4. **Umbrella Limit:**

* Most fraud reported cases have an umbrella limit of 0.
* We can note the same and for the non-fraudulent cases.

5. **Capital Gains and Losses:**

* No clear pattern.

6. **Incident Hour of the Day:**

* The data distribution varies.

7. **Number of Vehicles Involved:**

* Fraudulent cases are more likely to involve one vehicle.

8. **Bodily Injuries and Witnesses:**

* No clear pattern.

9. **Total Claim Amount, Injury Claim, Property Claim, Vehicle Claim:**

* No clear pattern.
* During the EDA process outliers have been identified and removed in the following columns - age', 'policy\_annual\_premium', 'umbrella\_limit', 'total\_claim\_amount' and 'property\_claim'. Outliers are extremely high or extremely low data point relative to the nearest data point and the rest of the neighboring co-existing values in a data graph. That`s why is very important to be identified and removed.
* Also the skewness (measure of the asymmetry of a distribution) has been identified and removed from the following columns - 'umbrella\_limit', 'total\_claim\_amount' and 'vehicle\_claim'.
* Encoding of the categorical columns had been performed.
* The correlation between the output ‘fraud\_reported’ and the other features been visualized and exanimated.

**4. Pre-processing Pipeline**

* After the data had been analysed and cleaned Feature Scaling using Standard Scalarization.
* Variance Inflation Factor VIF had been checked and any abnormal high values been identified. As a consequence of this the ‘total\_claim\_amount’ columns has been removed.
* Imbalance in the ‘fraud\_reported’ column data been identified and balancing had been performed.

**5. Building Machine Learning Models**

* The dataset had been split on training data (70%) and testing data (30%).
* As it was mentioned before there is a Binary classification problem.
* The model had been build and tested with different classification models.
* Cross Validation Score had been used to evaluate the model.
* After evaluating all classification models, the models with the highest accuracy score and cross validation score had been selected – Extra Tree and Random Forest Classifier.
* Hyperparameter Tuning had been performed to optimize the performance of the model and find the best combination of hyperparameters within predefined ranges.
* ROC-AUC curve has been build for both models.
* Finally the best on hyperparameter tuning the best model (Random Forest Classifier) had been selected and saved for future predictions of the fraudulent cases.

**6. Concluding Remarks**

Through the Explanatory Data Analysis and Data Visualization valuable insights of the data have been gained. This findings helped us to understand how the different features play role in identifying the fraudulent cases, which is very big issue in the Auto Insurance industry.