INSURANCE CLAIM FRAUD DETECTION MODEL

Submitted by:

Abhilekh Verma

Contents

ACKNOWLEDGMENT

I sincerely thanks to the Data Trained Faculty for the guidance. They have covered the topics like Machine Language, Python & SQL. Under their guidance I learned a lot about this project. their suggestions and directions have helped in the completion of this project. I had also taken help from YouTube & online videos.

INTRODUCTION

ABOUT INSURANCE FRAUD

Insurance fraud is any act committed to defraud an insurance process. It occurs when a claimant attempts to obtain some benefit or advantage they are not entitled to, or when an insurer knowingly denies some benefit that is due.

PROBLEM STATEMENT

Insurance fraud is a huge problem in the industry. It is difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem. In this project, you are provided a dataset which has the details of the insurance policy along with the customer details. It also has the details of the accident based on which the claims have been made. In this example, you will be working with some auto insurance data to demonstrate how you can create a predictive model that predicts if an insurance claim is fraudulent or not.

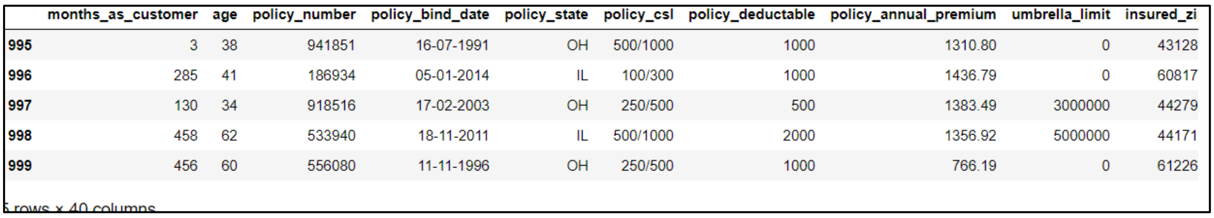
DATA ANALYSIS

COLUMNS IN THE DATAFRAME

- Months as customer
- age
- policy number
- policy bind date
- policy state
- policy csl
- policy deductible
- policy annual premium
- umbrella limit
- insured zip
- insured sex
- insured education level.
- insured occupation.
- Insured hobbies
- Insured relationship
- capital-gains.
- capital-loss
- incident date
- incident type
- collision type
- incident severity
- authorities contacted
- incident state
- incident city
- incident location
- incident hour of the day
- number of vehicles involved
- property damage
- bodily injuries
- witnesses
- police report available
- total claim amount
- injury claim
- property claim
- vehicle claim
- auto make
- auto model
- auto year
- fraud reported (Target)

Above all the columns are independent variable except ‘fraud reported’ column because it is our target column.

sample



	months_as_customer	age	policy_number	policy_bind_date	policy_state	policy_csl	policy_deductable	policy_annual_premium	umbrella_limit	insured_zi
995	3	38	941851	16-07-1991	OH	500/1000	1000	1310.80	0	43128
996	285	41	186934	05-01-2014	IL	100/300	1000	1436.79	0	60817
997	130	34	918516	17-02-2003	OH	250/500	500	1383.49	3000000	44279
998	458	62	533940	18-11-2011	IL	500/1000	2000	1356.92	5000000	44171
999	456	60	556080	11-11-1996	OH	250/500	1000	766.19	0	61226

5 rows × 40 columns

OBJECTIVE

- Our main objective predicts if an insurance claim is fraudulent or not.
- All the parameters will be analysed through Machine Learning algorithms like Logistic Regression, AdaBoost Classifier, Random Forest Classifier, Decision Tree Classifier, Support Vector Classifier etc which will help to predict.

DATA DESCRIPTION

- The source of data is taken from GitHub.
- Data(https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/Automobile_insurance_fraud.csv)

METHEDOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learnings techniques to determine insurance claim is fraudulent or not.

METRIC USAGE

a. Logistic Regression.
b. AdaBoost Classifier.
c. Random Forest Classifier.
d. Gaussian NB
e. Decision Tree Classifier.
f. Support Vector Classifier

SYSTEM REQUIREMENTS

HARDWARE and SOFTWARE REQUIREMENTS and TOOLS USED

a) Hardware Requirement:
   - i. Intel core i5
   - ii. 8 GB Ram

b) Software Requirement:
   - i. Python 3.x with packages:
     1. Pandas: Data analysis and manipulation tool
     2. NumPy: Provide support for mathematical functions, random number etc.
     3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
     4. Seaborn: is a library mostly used for statistical plotting in python.
     5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

APPROACH

IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES

- Logistic Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
- AdaBoost Classifier: s a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.
- Random Forest Classifier: The Random Forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.
- Gaussian NB: Gaussian NB is based on the Naive Bayes theorem with the assumption of conditional independence between every pair of features given the label of the target class.
- Decision Tree Classifier: A decision tree is a flowchart-like tree structure in which the internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. A Decision Tree consists of Nodes: Test for the value of a certain attribute.
- Support Vector Classifier: Support Vector Machine is a discriminative classifier that is formally designed by a separative hyperplane. It is a representation of examples as points in space that are mapped so that the points of different categories are separated by a gap as wide as possible. In addition to this, an SVM can also perform non-linear classification.
- Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
- Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
- ZScore: Z-score is also known as standard score gives us an idea of how far a data point is from the mean.
- Label Encoder: Label Encoding refers to converting the labels into numeric form.
- IMB Learn Anaconda: imbalanced-learn is a python package offering a number of re-sampling techniques commonly used in datasets showing strong between-class imbalance. It is compatible with scikit-learn and is part of scikit-learn-contra projects.
- Standard Scaling: Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance). It standardizes features by subtracting the mean value from the feature and then dividing the result by feature standard deviation.
- AUC-ROC CURVE: The AUC-ROC curve tells us visualize how well our machine learning classifier is carrying out. It is one of the most significant evaluation metrics for examining any classification model’s performance. It is also called as AUROC (Area Under the Receiver Operating Characteristics).

TESTING OF IDENTIFIED APPROACH(Algorithms)

a. Train Test Split
b. Label Encoding
c. STANDARD SCLAER (SMOTE)
d. IMB Learn
e. Logistic Regression
f. AdaBoost Classifier
g. Gaussian NB
h. Decision Tree classifier
i. Support Vector Classifier
j. Grid Search CV
k. Cross Validation
l. AUC ROC CURVE

KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

- Analysed data for any unique values.
- Analysed data for distribution.
- Compared between two columns.
- Checked and removed outliers through zscore method.
- Removed skewness present in the dataset.
- Done Oversampling.
- Cross validate the accuracy score from overfitting.
- Done AUC ROC score for better understanding.
- Hyper Parameter tuning using Grid Search CV.

CONCLUSION

As our conclusion we can say that , after checking accuracy score, cross validation, Ensemble Techniques, and checking AUC score we declare Decision Tree Classifier is predicting approx. 78.77% accuracy is best suited model for our purpose of predicting if an insurance claim is fraudulent or not.

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc. It also helped me to learn how to apply various model techniques on data and enable predications.

REFERENCES

- Data trained course videos.
- Google Search.
- YouTube.
- GitHub.
- UCI Machine learning repository.