Udacity Machine Learning Nanodegree
Fraud Detection on Insurance Claims

Proposal

Venghatesh S
25th Nov 2018

Domain Background

Insurance fraud is any act committed to defraud an insurance process. False insurance claims are insurance claims filed with an intent to defraud the insurance provider. Insurance fraud is deliberately undetectable, unlike visible crimes such as robbery or murder.

The Coalition Against Insurance Fraud estimates that in 2006 a total of about $80 billion was lost in the United States due to insurance fraud. According to estimates by the Insurance Information Institute, insurance fraud accounts for about 10 percent of the property/casualty insurance industry's incurred losses and loss adjustment expenses. The National Health Care Anti-Fraud Association estimates that 3% of the health care industry's expenditures in the United States are due to fraudulent activities, amounting to a cost of about $51 billion. Other estimates attribute as much as 10% of the total healthcare spending in the United States to fraud—about $115 billion annually. Another study of all types of fraud committed in the United States insurance institutions (property-and-casualty, business liability, healthcare, social security, etc.)put the true cost at 33% to 38% of the total cash flow through the system. In the United Kingdom, the Insurance Fraud Bureau estimates that the loss due to insurance fraud in the United Kingdom is about £1.5 billion ($3.08 billion), causing a 5% increase in insurance premiums. The Insurance Bureau of Canada estimates that personal injury fraud in Canada costs about C$500 million annually. India forensic Center of Studies estimates that Insurance frauds in India costs about $6.25 billion annually.

Hence it is highly imperative for the insurance industry to develop solid capabilities that can help identify potential frauds with a high degree of accuracy, so that other claims can be cleared rapidly while identified cases can be scrutinized in detail.

Citations :

https://www.insurancenexus.com/fraud/role-data-and-analytics-insurance-fraud-detection

https://www.researchgate.net/publication/291833022_Analytics_for_Insurance_Fraud_Detection_An_Empirical_Study

https://acadpubl.eu/jsi/2017-116-13-22/articles/21/80.pdf

Dataset could be found in the below location
https://raw.githubusercontent.com/jodb/sparkTestData/master/insurance_claims.csv

In this capstone project we will look at creating a predictive model that predicts whether a auto insurance claim is fraudulent or not.

Problem Statement

The current approach for Auto insurance fraud detection is based on a Heuristic approach around fraud indicators. Various scenario rules would be framed and these rules would determine along with the value of the claim to understand whether a claim needs to be sent for

Proposal

Venghatesh S
25th Nov 2018

investigation. The challenge with this approach is that it is heavily manual driven and has the following limitations

- Operates with a limited set of known parameters based on heuristic knowledge
- Not data driven
- Periodic manual re-calibration based on investigation outputs

This has led to auto insurance companies to explore ways of using machine learning algorithm which can look at the data patterns without judgement on relevance on data elements. Based on identified frauds, machine learning algorithms can develop a model through a variety of algorithmic techniques.

We will run a handful of machine learning algorithms to understand which features are relevant for fraud detection after doing data processing and transformation.

This is a **binary classification problem** since the end goal is to understand whether the claim is fraudulent or not.

Dataset and Inputs

Introduction to Datasets

Number of Claims	1000
Number of attributes	400
Categorical attributes	21
Normal Claims	753
Fraudulent Claims	247
Fraudulent Incident Rate	32.8%

Detailed Description

The number of input variables is 39 and output variable 1 is

months_as_customer int64	How long the customer has been with the Insurance company ?
age int64	Age of the Policy holder
policy_number int64	Policy number
policy_bind_date string	Effective date of the policy
policy_state string	US State where the insurance policy is taken
policy_csl string	Combined Single limit of the policy
policy_deductable int64	Amount to be paid out by the policy holder from his pocket

Udacity Machine Learning Nanodegree

Fraud Detection on Insurance Claims

Proposal

Venghatesh S

25th Nov 2018

policy_annual_premium float64	Annual premium amount to be borned by the policy holder
umbrella_limit int64	Limit for extra liability insurance.
insured_zip int64	Pincode of the policy holder
insured_sex string	Gender of the policy holder
insured_education_level string	Highest education level of the policy holder
insured_occupation string	Job profile
insured_hobbies string	Hobby of the policy holder
insured_relationship string	Relationship between policy holder and Beneficiary
capital-gains int64	Capital Gains
capital-loss int64	Capital Loss
incident_date string	Accident Date
incident_type string	Type of incident (Eg : Vehicle theft, single vehicle collison etc)
collision_type string	Type of Collison (Side, Rear, Front)
incident_severity string	Minor Damage, Total Loss, Major Damage
authorities_contacted string	Authorities approached to validate the claim
incident_state string	State where the incident occurred
incident_city string	City where the incident occurred
incident_location string	Exact spot where the incident occurrent
incident_hour_of_the_day int64	Time of the accident
number_of_vehicles_involved int64	Self Explantory
property_damage string	Any property damage (Yes or No or unknown)
bodily_injuries int64	No of people with body injury
witnesses int64	Number of people witness to the accident
police_report_available string	Avalability of police report (Yes or no)
total_claim_amount int64	Total Claim

Udacity Machine Learning Nanodegree

Fraud Detection on Insurance Claims

Proposal

Venghatesh S

25th Nov 2018

injury_claim int64	Injury Claim Amount
property_claim int64	Property damage Claim Amount
vehicle_claim int64	Vehicle damage claim amount
auto_make string	Auto Manufacturer
auto_model string	Model name of the Auto
auto_year int64	Auto production year
fraud_reported string	Fraud (Yes or No). This is the output variable
_c39 float64	This variable should be dropped

- We have several categorical variables (String) datatypes in the dataset . Some Statistical models which uses calculation of distances (Euclidean or Mahlanobis or other measures) can handle only numerical attributes. So, these categorical attributes should be converted to numerical attributes through encoding.
- Some of the features in the dataset have got lot of distinct values and they won't be useful in prediction . Having them will increase dimensionality of the dataset with no considerable improvement in prediction due to them. We should look at removing them.

Some of the data may be wrong . Eg : age of the driver may be marked as 2, which is not practically possible. An analysis of the dataset should be done to check for any anomalies

Dataset could be found in the below location
https://raw.githubusercontent.com/jodb/sparkTestData/master/insurance_claims.csv

Solution Statement

The solution would be for the model to predict will be predictions of whether a auto insurance claim is fraud or not. This is a classification problem and most common solution to these problems are to use SVM,Decision tree,Random Forest, Gradient Boosting even though all of them can be used for regression problems also. Fraud detection patterns are complex and the current dataset contains only around 1000 records. SVM is generally used when the dataset size is very less and features are high. This is for the precise reason SVM is chosen as one of the model. Disadvantage of SVM is that they tend to take lot of time during training. Decision trees are easy to visualise, takes less time to train. The disadvantage is that they tend to overfit . The overfit could be reduced by simply averaging out the predictors like in Random Forest or by repeatedly train trees on the residuals of the previous predictors like GBM. This is for the reason Decision Tree, Random Forest and GBM are chosen as other model solutions along with SVM.

Benchmark Model

There is no known universal benchmark available for this model.
Bench Mark selection process :

1) Run the model on dataset for all the models (SVM, Decision Tree, Random Forest, Gradient Boosting with out of the box parameters.
2) Choose the model with highest accuracy score as the bench mark. Based on gut feeling GBM might perform better and that would be taken as the bench mark model.

Udacity Machine Learning Nanodegree
Fraud Detection on Insurance Claims

Proposal

Venghatesh S
25th Nov 2018

3) This model will be further fine tuning the features and parameter selection so that it can beat the bench mark.

Evaluation Metrics

Accuracy is a common metric for binary classifiers; it takes into account both true positives and true negatives with equal weight.

Accuracy alone doesn't tell the full story when you're working with a **class-imbalanced data set**, like this one, where there is a significant disparity between the number of positive and negative labels. Two better metrics for evaluating class-imbalanced problems: precision and recall.

To fully evaluate the effectiveness of a model, you must examine **both** precision and recall. Unfortunately, precision and recall are often in tension. That is, improving precision typically reduces recall and vice versa.

The best metric hence would be something which uses both recall and accuracy. That is called as the F1 Score.
The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

Project Design

Programming language :

- Python 3
- Sci-kit learn
- Seaborn and matplotlib

Design steps

1. Read data file
2. Data Visualization & Data Pre-processing
    a. Drop irrelevant features (features with distinct values and we are sure that they don't play any role in model based on domain expertise. prediction (e.g.: Policy number)
    b. Convert Categorical features in to numerical features
    c. Remove skewness
    d. Fix any data issues
    e. Look for co-relation between different features.
3. Model evaluation
    a. Evaluate model based on accuracy score and F-score for SVM, Decision Tree, Random Forest, Gradient Boost.
4. Choose model based on best accuracy and F-score
5. Have the score obtained from above as benchmark
6. Fine tune the model by using feature selection by looking at important features obtained using Gridsearch.
7. Fine tune the model further by tuning various parameters of the model and beat the accuracy and F-scores got during benchmark.

Udacity Machine Learning Nanodegree
Fraud Detection on Insurance Claims

Proposal

Venghatesh S
25th Nov 2018

References :

https://en.wikipedia.org/wiki/Insurance_fraud

https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

https://en.wikipedia.org/wiki/Support_vector_machine

https://en.wikipedia.org/wiki/Decision_tree