Person of Interest Detection Model

Rahul Gupta, MS in Software Engineering

Arizona State University, United States

Abstract

The aim of this report is to provide the detailed overview of the project 'Person of Interest Detection using Machine Learning' completed in October'2017. This self-initiated project was done as a part of the two months course 'Intro to Machine Learning' provided on Udacity. This course provided end-to-end process of investigating data through the lens of machine learning, and helped in applying the learning in real-world data set. In this project, famous Enron Dataset has been used to detect possible Person of Interest (POI). This POI Model was created using several classifiers with different classifier scores. This report discusses the learning and application process for the project in detail.

Keywords: *Person of Interest, Machine Learning, Enron, python, Decision tree*

Abbreviation Used*: Machine Learning – ML, Person of Interest (POI)*

Machine Learning is a first-class ticket to the era of most exciting new technology today. As data sources proliferate along with the computing power to process them, heading towards data is one of the most sensible ways to quickly gain some insights and make predictions. This course and the project has provided a similar insight to wrestle all the raw data into refined trends and predictions. The famous dataset of Enron, company which went into bankruptcy, has been used in training of Machine Learning algorithm. The project creates the predictive model to identify Enron employees who may have committed fraud based on the public Enron and email dataset.

This report elaborates on the datasets used, classifier to create the model with multiple metric estimators, the programing language support taken while designing, implementing and validating the model. It also brings light on how the model can be improved by incorporating other strategies to the model. The report contains the project results demonstrating the effectiveness of the model.

Project Overview

Dataset: In 2000, Enron was one of the largest energy-trading companies in United States. By 2002, due to corporate fraud, company had collapsed into bankruptcy. It was one of the biggest shakedowns of corporate world. During federal Investigation, in 2003, significant amount of typically confidential information entered into public record, including tens of thousands of emails and detailed financial data of top executives. Now, Enron E-mail corpus is available online and is valuable to various computer scientists because it is a rich example of how people in an organisation use email for several real-time plans. For a field as diverse as Machine Learning, this dataset would be perfect for research studies. Several softwares like fraud detection or workplace behavioural patterns over e-mail are benefited by this dataset. In this project, we are interested in labelling every person in the dataset into either a POI or a non-POI.

Machine Learning (ML): Machine learning is a field that gives computers the ability to learn without being explicitly programmed and makes programming paradigms open. In this project Machine learning concepts have been applied on Enron dataset to make smarter prediction model which can detect fraud based on its trained model. A ML project might not be linear in nature, but it can be broken down in numerous steps:

- Problem Definition
- Data Preparation and summarizing
- Algorithm Evaluation
- Making Prediction
- Result improvement
- Model testing
- Result presentation

Various languages can implement the above steps like python, R, etc. In this project, python is used as a programming tool over Machine Learning Platform. Several Data Manipulation and Machine Learning concepts are used:

- Feature selection and extraction: Features or variables in the dataset to which model is most affected was selected or prioritised to take part in training of model.

- Outlier Removal: Since, outliers are the rare cases. So to improve quality of the prediction model, outliers were removed.
- Cross-Validation: It is used to validate the model. This technique estimates the performance and stability of the model.
- Classification: It is a systematic approach to build classification models in which data is efficiently organised into categories. It makes data easier to retrieve. In this project, out of various known classifiers, decision tree classifier was used.
- Accuracy Score: It is the measure of accuracy of the model. It estimates the performance of the model to identify the new data.
- Precision and Recall: It is another metric which is used to measure how well an information retrieval system reveals the information. Both are measure relevance.

Programming Language: The project has used Python as the programming language support. Since, Python has numerous packages for Machine Learning and other computations like numpy, sklearn, etc., it provided easier user-friendly interface to develop the stated model. Online structured documentation of python was very helpful in developing the project. Python version 3.x has been used as a basic programming tool to create, validate, train and test the predictive model. During the improvement phase of the model, python modules and packages helped in fast training of the models.

Implementation

This model was implemented using the resources (predefined python files, data-sets, etc.) provided by the above mentioned coursework 'Intro to Machine Learning'. Decision tree was used as MI algorithm in the model. After complete analysis, results were recorded. Here is the complete description of all the concepts, language, algorithms and metrics implemented in the project.

Programming Platform:

- Installation of Python 3.x and various other packages like sklearn, numpy etc.
- Importing of predefined python packages and modules like sklearn, etc. for implementation.

Data Pre-processing:

- Outlier Removal was done by visualizing the data in Python module matplotlib. In Enron data, clear outlier was feature-total. It was removed from the dataset for further analysis.
- Feature selection and extraction have been used on the data-set.
- To increase the performance of the model, new features were created from the dataset. These new feature were created using feature extraction and then were added in the dataset.

Algorithm Selection:

The performance of model depends on selection of algorithm and selection of algorithm depends on the type of problem. Since, the objective of the model was to classify and find out whether someone is POI or Non-POI, Decision Tree algorithm seem to be an appropriate choice. Decision Tree organises a series of test questions and conditions in a tree structure. The root and internal nodes contain attribute test

conditions. After construction of decision tree, classifying a test record is straightforward. Starting from the root node, we apply the test condition to the record and follow the appropriate branch based on the outcome of the test.

Decision Tree has various parameters which were manually tuned to improve performance of the model.

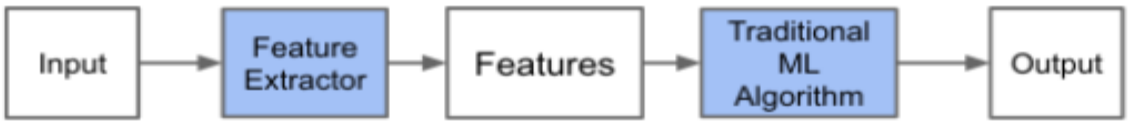
Performance Validation:

Validation is required to estimate the performance of any predictive model. Hence, K-Fold cross validation technique was used in the model on the Enron Data-Set. Cross-Validation was done on basis of selected features by splitting the training dataset into two datasets. One dataset (around 80% of total dataset) was used to train the model and other dataset (20% of total dataset) validated the model.

Estimation and Improvement analysis:

In this model, estimation of performance was determined by the accuracy score, precision and recall. They were calculated and recorded each time the classifier was trained. Improvement in model was done by changing the feature count or classifier parameter of the decision tree classifier which, in theory, should give better results. Accuracy score describes the model's ability to estimate positively. Precision means that an algorithm returned substantially more relevant results than irrelevant ones, while high recall means that an algorithm returned most of the relevant results.

ML Data Flow



Traditional Machine Learning Flow

The above figure is the general overview flow diagram of how the machine learning flow occurs to create any kind of predicting model. First, feature Extraction is done on pre-processed input data to extract various required features. After feature extraction, those features along with data is fed into ML algorithm, in this case Decision Tree Algorithm. After training and validation, the estimating measures for the data were taken to provide some improvements.

Practically, Traditional ML algorithm and Output are in loop. To improve the results of the model, after measuring outputs, flow goes back to ML algorithm, where after several changes Output is taken again. This process goes on till there is further scope of improvement in the designed model. Likewise, in this model, decision tree classifier have been trained multiple times with different number of features and different parametric values of the classifiers used.

Results

The project implementation was done using decision tree classifier. Results have been recorded before tuning and after tuning of decision tree. Tuning of the classifier was done by initialising min_samples_split of decision tree to 5. Table 1 depicts the accuracy score of the model, and table 2 shows the precision and recall measures of the model.

Accuracy Score:

	Decision Tree before tuning	Decision tree after tuning
All Features	89.58	89.58
Top 5 contributing features	92.86	92.43

(Table 1)

Precision and Recall:

	Precision	Recall
All Features	0.5	0.4
Top 5 contributing features	0.75	1.0

(Table 2)

The results show the ability of the model to estimate the presence of fraud in a particular dataset. Both the metric measures (Accuracy and Precision & Recall) clearly show that significant amount of improvement can be seen in the model if feature selection is done properly.

Interpretation of Results:

We can see that, the model is providing favourable results. A fraud identification model can be created and improved extensively if understood properly. As per the shown results in Table 1, it can be improved -

- By using certain defined parameters affecting the algorithm or classifier, or
- By selecting right features with which model is properly attached.

And, as per Table 2, it can be interpreted that there is a significant amount of improvement in the model. Increase in Recall shows that the five features taken in the model are pretty relevant and increased precision value show relevance in the model.

Learning from Project:

The knowledge imparted by Project is many folds. Following are the major learnings from this project:

- Ability to deal with real time dataset.
- How to validate the machine learning results using dataset.
- Evaluation of a machine learning result using quantitative metrics
- Creation, selection and transformation of features to obtain better model accuracy.
- Performance comparison of machine learning algorithms
- Tuning of machine learning algorithms for maximum performance and estimations.
- Ability to communicate your machine learning algorithm results clearly.

- Understanding of python libraries related to Machine Learning and data manipulation.
- Understanding the flow of machine learning algorithm.
- Detail analysis on Decision Tree Algorithm.

Conclusion

This project depicts a naïve understanding of analysis for classifying Enron employees. The results should not be taken too seriously, since more advanced models should be used. The purpose of this project is to obtain understanding of several concepts of ML, python, and data manipulation.

Further Improvements

This project is just for intermediate understanding of how the process flow from data manipulation and pre-processing to Prediction model. There are various other ML algorithms which can also be used to test and improve the results. More Improvements can be seen by implementing the ensemble concepts where multiple classifiers are used and their average is taken to improve the prediction model.

Today, techniques like neural network, deep learning, etc. are used to create better models which can provide better results and work fairly in much broader sense.