**Machine Learning Engineer Nanodegree**

**Capstone Project**

Santhana Sankaramurthy  
November 12, 2017

**I. Definition**

*(approx. 1-2 pages)*

**Project Overview**

This project is intended to build a machine learning algorithm to effectively detect fraud in credit card transactions. This problem space has been extensively researched and implemented in industry over the years as is evident from the work of Chan and Stolfo – Towards Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection available at http://www.aaai.org/Papers/KDD/1998/KDD98-026.pdf

Industry estimates indicate that 0.1% of credit card transactions globally are fraudulent [Source: Wikipedia]. Since typical fees charged by credit card companies and online payment providers are between 2 and 4% of total transaction values, even a small improvement in fraud detection contributes significantly to improved profitability and competitiveness in the marketplace.

Increasing fraud ultimately results in increasing costs for legitimate card holders, since companies need to recoup these costs in some way. Since credit providers bear the cost of fraud, credit providers that do better at fraud detection are able to charge lesser fees to card holders and be more competitive in the marketplace, both to merchants that use these card providers as well as individual card holders.

More broadly speaking, these algorithms can be used for other imbalanced data sets that occur widely in nature and business, some of which are even more imbalanced e.g. factory production error rates (about 0.1%), server failure rates, cancer detection (0.45%), detecting oil slicks from satellite ocean images, etc.

We will use the dataset for credit card fraud available from Kaggle. This is available at the link - https://www.kaggle.com/dalpozz/creditcardfraud. This consists of real data from European cardholders in 2013. The data consists of 30 total features [i.e. 28 principal components identified through PCA, transaction amount, time [elapsed since first transaction] along with the Class for each transaction - 1 indicating fraud and 0 indicating normal transaction. There are 492 frauds out of 284,807 total transactions in the data provided, hence this is a highly imbalanced dataset with the positive class amounting to only 0.172% of all transactions.

In this section, look to provide a high-level overview of the project in layman’s terms. Questions to ask yourself when writing this section:

* *Has an overview of the project been provided, such as the problem domain, project origin, and related datasets or input data?*
* *Has enough background information been given so that an uninformed reader would understand the problem domain and following problem statement?*

**Problem Statement**

In this project, we will create an algorithm to identify which transactions are fraudulent from the given transaction set. All the transactions in the dataset provided have been labeled as being in one of two potential classes i.e. fraudulent (Class 1) or legitimate (Class 0), we will use supervised learning and specifically, classification algorithms to determine which class each transaction belongs to.

As a first step, we will examine the data set and identify any characteristics that can potentially help us define the algorithm. At this stage, we will also graph the data to examine it visually and identify any patterns. None of the features are categorical, hence none need to be converted to numerical ones before we can feed them to the classifier.

Then, we will pre-process the data to scale the features and ensure that none of them dominate simply due to their higher numerical values.

Examples of classification algorithms that can be used are Logistic Regression, Decision Trees, Naïve Bayes, Random Forest, Simple Vector Machines and Boosting (Gradient Boosting, Ada Boost, XGBoost ) As the data is highly imbalanced, we will consider only algorithms that will work effectively with such datasets – these are Decision Trees, Random Forest and Boosting methods. Once we do the initial evaluation of the algorithms based on relative performance (based on metrics that are defined in the next section), we can identify which one is suitable for further tuning. This can be further tuned using hyper parameter optimization to improve performance further.

In this section, you will want to clearly define the problem that you are trying to solve, including the strategy (outline of tasks) you will use to achieve the desired solution. You should also thoroughly discuss what the intended solution will be for this problem. Questions to ask yourself when writing this section:

* *Is the problem statement clearly defined? Will the reader understand what you are expecting to solve?*
* *Have you thoroughly discussed how you will attempt to solve the problem?*
* *Is an anticipated solution clearly defined? Will the reader understand what results you are looking for?*

**Metrics**

In this section, you will need to clearly define the metrics or calculations you will use to measure performance of a model or result in your project. These calculations and metrics should be justified based on the characteristics of the problem and problem domain. Questions to ask yourself when writing this section:

* *Are the metrics you’ve chosen to measure the performance of your models clearly discussed and defined?*
* *Have you provided reasonable justification for the metrics chosen based on the problem and solution?*

We will evaluate algorithm performance by using the following two metrics –

* Primary Metric: AUPRC (Area under the precision-recall curve) – we will use this as opposed to the AUC since the dataset is highly imbalanced and the simple AUC measure (which measures the area under the ROC curve will tend to show a high value even if the classifier is not performing particularly well). AUPRC is provided by sklearn as the average\_precision score
* Since we are interested in ensuring that we identify as many of the fraudulent transactions as possible (preferably all), we will use the recall score. This is the number of fraudulent transactions correctly identified divided by the actual number of fraudulent transactions in the test dataset. The higher the recall accuracy, the better the classifier is able to identify all fraudulent transactions.
* The ultimate objective for the credit card provider is to minimize the loss due to fraudulent transactions. In this context, a fraudulent transaction for $ 2,000 obviously costs the credit provider way more than a fraudulent transaction that costs $ 20. Measuring this will mean determining the total loss for all misclassified transactions. This can be defined as the sum of the following
  + For False Negative transactions, the loss is the transaction amount
  + For False positive transactions, the customer will have to prove that the transaction is legitimate and hence the loss can be viewed as a notional amount (to cover the cost of processing + potential annoyance to the customer)

**II. Analysis**

*(approx. 2-4 pages)*

**Data Exploration**

There are 284,807 data points with 31 variables each. Of these, 492 transactions have been identified as fraudulent.

There are 30 features available in the dataset – ‘Amount’ indicating transaction amount, ‘Time’ indicating the time elapsed since the first transaction and V1, V2…V28 – 28 other numerical features that are output by a PCA on the raw data. There are no categorical features and no missing values.

There is also the ‘Class’ feature which has a value of 1 for fraud and 0 for normal transactions. A box plot of the feature information is provided below.

<<Insert boxplot>>

A simple scatter plot with Amount on the vertical axis and time on the horizontal axes is provided below – this indicates that there is no visible variation in fraud across time, hence the ‘Time’ feature can be safely dropped from the data being fed into the classifier.

<<Insert graph>>

The highest value of Amount for the fraud transactions is $ 2,125.87. Hence all transactions above this value can probably be safely dropped. However, since the cost of an error here i.e. missing a fraudulent transaction beyond $ 2,126 will be quite expensive. In addition, since this type of a rule will not generalize well beyond this particular dataset, we will keep the entire dataset as is.

In this section, you will be expected to analyze the data you are using for the problem. This data can either be in the form of a dataset (or datasets), input data (or input files), or even an environment. The type of data should be thoroughly described and, if possible, have basic statistics and information presented (such as discussion of input features or defining characteristics about the input or environment). Any abnormalities or interesting qualities about the data that may need to be addressed have been identified (such as features that need to be transformed or the possibility of outliers). Questions to ask yourself when writing this section:

* *If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader?*
* *If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed?*
* *If a dataset is* ***not*** *present for this problem, has discussion been made about the input space or input data for your problem?*
* *Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)*

**Exploratory Visualization**

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

* *Have you visualized a relevant characteristic or feature about the dataset or input data?*
* *Is the visualization thoroughly analyzed and discussed?*
* *If a plot is provided, are the axes, title, and datum clearly defined?*

**Algorithms and Techniques**

**Example dependent cost classification**

**Sklearn weights**

**How do data-based techniques fit into the overall flow? i.e. if you resample the data, it will not be comparable to results with the base data**

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

* *Are the algorithms you will use, including any default variables/parameters in the project clearly defined?*
* *Are the techniques to be used thoroughly discussed and justified?*
* *Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?*

**We will use the following basic classification algorithms**

Naïve Bayes – from a study of the research, it is clear that this is unlikely to perform well since the algorithm biases prediction in favor of the majority class i.e. [No Fraud]. Hence, we will need to balance the dataset prior to using this model.

Logistic Regression or other scikit-learn algorithms with modified class weights. For modified sampling methods

**Benchmark**

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

* *Has some result or value been provided that acts as a benchmark for measuring performance?*
* *Is it clear how this result or value was obtained (whether by data or by hypothesis)?*

**III. Methodology**

*(approx. 3-5 pages)*

**Data Preprocessing**

On viewing a boxplot of the data, it is clear that a few features are at a different scale compared to the rest

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

* *If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?*
* *Based on the* ***Data Exploration*** *section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?*
* *If no preprocessing is needed, has it been made clear why?*

**Implementation**

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

* *Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?*
* *Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?*
* *Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

**Refinement**

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

* *Has an initial solution been found and clearly reported?*
* *Is the process of improvement clearly documented, such as what techniques were used?*
* *Are intermediate and final solutions clearly reported as the process is improved?*

**IV. Results**

*(approx. 2-3 pages)*

**Model Evaluation and Validation**

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model’s solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

* *Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?*
* *Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?*
* *Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?*
* *Can results found from the model be trusted?*

**Justification**

In this section, your model’s final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

* *Are the final results found stronger than the benchmark result reported earlier?*
* *Have you thoroughly analyzed and discussed the final solution?*
* *Is the final solution significant enough to have solved the problem?*

**V. Conclusion**

*(approx. 1-2 pages)*

**Free-Form Visualization**

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

* *Have you visualized a relevant or important quality about the problem, dataset, input data, or results?*
* *Is the visualization thoroughly analyzed and discussed?*
* *If a plot is provided, are the axes, title, and datum clearly defined?*

**Reflection**

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

* *Have you thoroughly summarized the entire process you used for this project?*
* *Were there any interesting aspects of the project?*
* *Were there any difficult aspects of the project?*
* *Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?*

**Improvement**

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

* *Are there further improvements that could be made on the algorithms or techniques you used in this project?*
* *Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?*
* *If you used your final solution as the new benchmark, do you think an even better solution exists?*

**Before submitting, ask yourself. . .**

* Does the project report you’ve written follow a well-organized structure similar to that of the project template?
* Is each section (particularly **Analysis** and **Methodology**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
* Would the intended audience of your project be able to understand your analysis, methods, and results?
* Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?
* Are all the resources used for this project correctly cited and referenced?
* Is the code that implements your solution easily readable and properly commented?
* Does the code execute without error and produce results similar to those reported?