# **Datascience Nanodegree**

## **Capstone Project**

Roald Brønstad  
November 26, 2018

## **I. Definition**

### **Project Overview**

Predicting stock prices and market direction can surely be considered the holy grail within finance. Algorithmic trading is replacing humans and the finance sector is exploring how new technology can be utilized every day.

Traditionally the finance sector has relied on econometric tools such as linear regression to try and predict the market, today machine learning and deep learning networks are taking over the task.

The purpose of this project is to investigate the possibility of predicting tomorrows stock market direction based on historical data as a binary classification, up or down. Daily prices for the Norwegian stock index OSEBX is available online with variables such as open, high, low, closing price and volume. I have also expand the data-set with technical indicators such as momentum, relative strength index, moving average etc. The difference between closing price today and closing price tomorrow (prediction) will be either positive or negative, indicating direction.

The project will experiment with several machine learning algorithms as well as artificial neural network to analyze the possibility of predicting tomorrow direction

### **Problem Statement**

A successful model will be able to predict whether the market will move up or down tomorrow better than random guess. It should output a 1 for predicting upwards movement and 0 for downward movement.

An important step before building this model will be to create the data-set. The data is a time-series and in order to use machine learning models in a useful way, technical indicators must be derived from the time series observations. The target label (at time t) will be defined as 1 if closing price for (t+1) - t is positive and 0 if its negative.

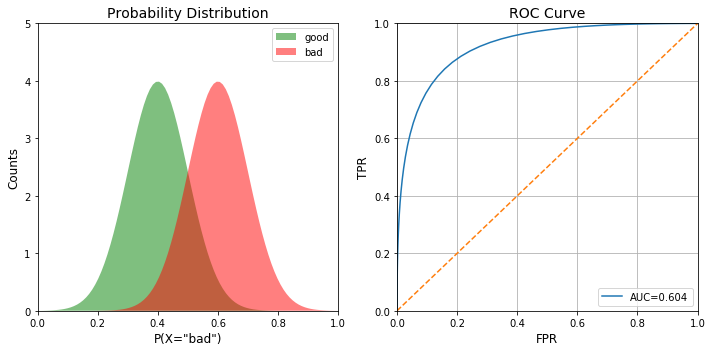
The data-set will be split into a training set and test set by a ratio of 70 - 30% and normalized to a range between 0 and 1.

I will train a simple logistic regression model to serve as a benchmark model to compare against more advanced and relevant machine learning models such as Random forest, Naive Bayes, Support vector machines and Adaboost. The models will be trained and hyper tuned by exhaustive grid-search and compared to each other and the benchmark model. I will also build a Recurrent neural network which has been developed especially to work well on sequence problems.

**Metrics**

I have used ROC curve and area under curve as an evaluation metric for this project.

A roc curve is derived from a confusion matrix of the test data (true/false positive and true/false negatives) A binary classifier will classify based on the distribution of probabilities. (see picture below) A threshold needs to be set unless the model can predict 100%. The ROC curve shows true positive rate (true positive / all positive) against a false positive rate (false positive / all negative) plotted for all discriminating threshold in the distributions. The diagonal line shows pure guessing between two classes. The ROC curve for a model with good predictive power will lie close to the upper left corner. Far away from the diagonal line.



## **II. Analysis**

(approx. 2-4 pages)

### **Data Exploration**

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**

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?

### **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**

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?