**Computer Science**

**Mid term report on mini project**

On

**Stock Market Prediction**

Based On

**Machine Learning**

Done By

**SHIVAM UPADHYAY**

**(171500324)**

In Under Guidance Of

**Mr. Subhash Agrawal**

**(**Asstt. Professor)

Submitted To

****

**Department Of Computer Science And Engineering**

**Abstract**

In this report we analyse existing and new methods of stock market prediction. We take three diﬀerent approaches at the problem: Fundamental analysis, Technical Analysis, and the application of Machine Learning. We ﬁnd evidence in support of the weak form of the Eﬃcient Market Hypothesis, that the historic price does not contain useful information but out of sample data may be predictive. We show that Fundamental Analysis and Machine Learning could be used to guide an investor’s decisions. We demonstrate a common ﬂaw in Technical Analysis methodology and show that it produces limited useful information. Based on our ﬁndings, algorithmic trading programs are developed and simulated using Quantopian.

**Table of Contents**

## Table of Contents

1. Introduction
2. Process of Machine Learning Predictions
3. Housing Data Set
   * Understand the problem
   * Hypothesis Generation
   * Get Data
   * Data Exploration
   * Data Pre-Processing
   * Feature Engineering - Create 331 new features
   * Model Training - XGBoost, Neural Network, Lasso
   * Model Evaluation

**Introduction**

For freshers, projects are the best way to highlight their data science knowledge. In fact, not just freshers, up to mid-level experienced professionals can keep their resumes updated with new, interesting projects. After all, they don't come easy. It takes a lot of time to create a project which can truly showcase the depth and breadth of your knowledge.

I hope this project will help you gain much needed knowledge and help your resume get shortlisted faster. This project shows all the steps (from scratch) taken to solve a Machine Learning problem. For your understanding, I've taken a simple yet challenging data set where you can engineer features at your discretion as well.

We'll participate in a Kaggle competition and make our way up the leaderboard among ~ top 14% participants.

This project is most suitable for people who have a basic understanding of python and Machine Learning. Even if you are absolutely new to it, give it a try. And ask questions in Comments below. R users can refer to this equivalent R script and follow the explanation given below.

## Process of Machine Learning Predictions

“Keep tormenting data until it starts revealing its hidden secrets.” Yes, it can be done but there's a way around it. Making predictions using Machine Learning isn't just about grabbing the data and feeding it to algorithms. The algorithm might spit out some prediction but that's not what you are aiming for. The difference between good data science professionals and naive data science aspirants is that the former set follows this process religiously. The process is as follows:

1. **Understand the problem:** Before getting the data, we need to understand the problem we are trying to solve. If you know the domain, think of which factors could play an epic role in solving the problem. If you don't know the domain, read about it.

2. **Hypothesis Generation:** This is quite important, yet it is often forgotten. In simple words, hypothesis generation refers to creating a set of features which could influence the target variable given a confidence interval ( taken as 95% all the time). We can do this before looking at the data to avoid biased thoughts. This step often helps in creating new features.

3. **Get Data:** Now, we download the data and look at it. Determine which features are available and which aren't, how many features we generated in hypothesis generation hit the mark, and which ones could be created. Answering these questions will set us on the right track. 4.

**Data Exploration:** We can't determine everything by just looking at the data. We need to dig deeper. This step helps us understand the nature of variables (skewed, missing, zero variance feature) so that they can be treated properly. It involves creating charts, graphs (univariate and bivariate analysis), and cross-tables to understand the behavior of features.

5. \*Data Preprocessing: \*Here, we impute missing values and clean string variables (remove space, irregular tabs, data time format) and anything that shouldn't be there. This step is usually followed along with the data exploration stage.

6. **Feature Engineering:** Now, we create and add new features to the data set. Most of the ideas for these features come during the hypothesis generation stage.

7. **Model Training:** Using a suitable algorithm, we train the model on the given data set.

8. **Model Evaluation:** Once the model is trained, we evaluate the model's performance using a suitable error metric. Here, we also look for variable importance, i.e., which variables have proved to be significant in determining the target variable. And, accordingly we can shortlist the best variables and train the model again.

9. **Model Testing:** Finally, we test the model on the unseen data (test data) set.

We'll follow this process in the project to arrive at our final predictions. Let's get started.

## 1.Understand the problem

The data set for this project has been taken from Kaggle's [Housing Data Set](https://www.kaggle.com/c/house-prices-advanced-regression-techniques) Knowledge Competition. As mentioned above, the data set is simple. This project aims at predicting house prices (residential) in Ames, Iowa, USA. I believe this problem statement is quite self-explanatory and doesn't need more explanation. Hence, we move to the next step.

## 2. Hypothesis Generation

Well, this is going to be interesting. What factors can you think of right now which can influence house prices ? As you read this, I want you to write down your factors as well, then we can match them with the data set. Defining a hypothesis has two parts: Null Hypothesis (Ho) and Alternate Hypothesis(Ha). They can be understood as:

Ho - There exists no impact of a particular feature on the dependent variable. Ha - There exists a direct impact of a particular feature on the dependent variable.

Based on a decision criterion (say, 5% significance level), we always 'reject' or 'fail to reject' the null hypothesis in statistical parlance. Practically, while model building we look for probability (p) values. If p value < 0.05, we reject the null hypothesis. If p > 0.05, we fail to reject the null hypothesis. Some factors which I can think of that directly influence house prices are the following:

* Area of House
* How old is the house
* Location of the house
* How close/far is the market
* Connectivity of house location with transport
* How many floors does the house have
* What material is used in the construction
* Water /Electricity availability
* Play area / parks for kids (if any)
* If terrace is available
* If car parking is available
* If security is available

…keep thinking. I am sure you can come up with many more apart from these.

## 3. Get Data

You can [download the data](https://www.kaggle.com/c/house-prices-advanced-regression-techniques) and load it in your python IDE. Also, check the competition page where all the details about the data and variables are given. The data set consists of 81 explanatory variables. Yes, it's going to be one heck of a data exploration ride. But, we'll learn how to deal with so many variables. The target variable is SalePrice. As you can see the data set comprises numeric, categorical, and ordinal variables. Without further ado, let's start with hands-on coding.

## 4. Data Exploration

Data Exploration is the key to getting insights from data. Practitioners say a good data exploration strategy can solve even complicated problems in a few hours. A good data exploration strategy comprises the following:

1. **Univariate Analysis** - It is used to visualize one variable in one plot. Examples: histogram, density plot, etc.
2. **Bivariate Analysis** - It is used to visualize two variables (x and y axis) in one plot. Examples: bar chart, line chart, area chart, etc.
3. **Multivariate Analysis** - As the name suggests, it is used to visualize more than two variables at once. Examples: stacked bar chart, dodged bar chart, etc.
4. **Cross Tables** -They are used to compare the behavior of two categorical variables (used in pivot tables as well).

