**Chapter 2 – Predicting Economic Cycle based on CP and t-Bills**

1. **Introduction**

The underlying financial idea behind this study is that the spread on commercial paper, a short term form of corporate borrowing, and the US Treasury bill widens before recessions and contracts after and could be a useful predictor of real economic activity. This is similar to credit spreads widening before a stock market crash(eg in 1987 and 2008). There is considerable literature on this subject from a financial standpoint; our study in this project is to use sophisticated machine learning techniques to draw conclusions about this. The Economic Indicator that we are going to monitor is the ***USHPCI Index***.

There are 223 sets of observations, with features like US treasury rates across the curve, the short term Commercial Paper borrowing rates, their spreads. The target variables are the 3/6/9 month forward change in the USHPCI Index based on the above inputs.

The rest of this study is divided into the following subsections:

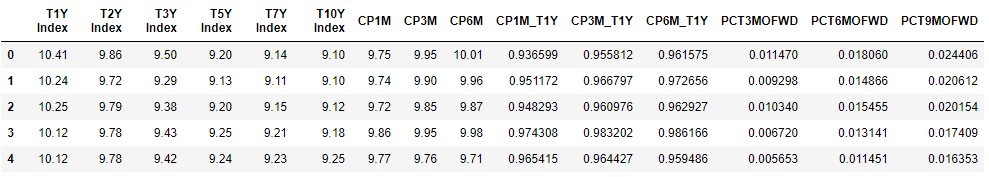
* Studying the dataset, Preprocessing, handling missing/0 values etc
* Exploratory data analysis to get an idea of driving factors, correlations etc
* Model fitting – Three models have been used: **Linear Regression**, **Regression Decision Tree** and **Support Vector Regression**(SVR)
* 10 fold Cross Validaton
* Random Forrest Regressor has also been implemented and yields the best result
* Conclusions – summary/findings

Target Variables: PCT3MOFWD, PCT6MOFWD, PCT9MOFWD

Each of the above three are separately modelled and fitted with the attributes. The results are studied for each and an inference derived.

1. **Exploratory Data Analysis & Preprocessing**

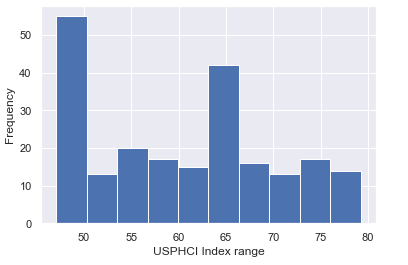
The following table shows a basic summary of the dataset:



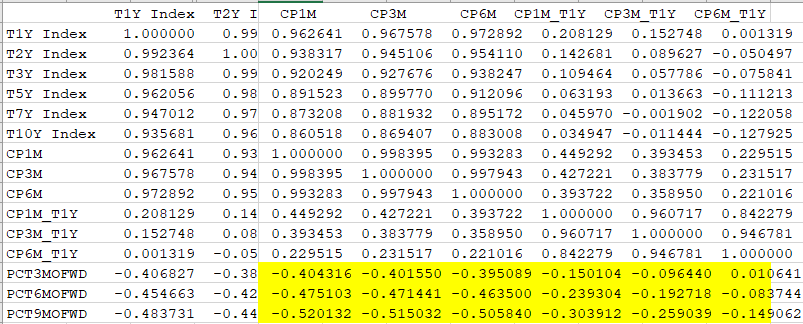
The basic characteristics of the features was found out using the describe function. As an example:

|  |
| --- |
| count 222.000000  mean 0.007092  std 0.004848  min -0.006811  25% 0.005567  50% 0.008272  75% 0.010206  max 0.020297 |
| Name: PCT3MOFWD, dtype: float64 |

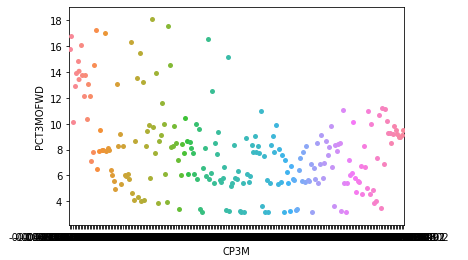
‘0’ values: PCT9MOFWD has one observation as 0. So, I first replaced it by NAN using np.nan then drop it since its only one row. I did not impute the data as there were no other 0/missing observations.

Histogram for the USHPCI Index

**Correlations**



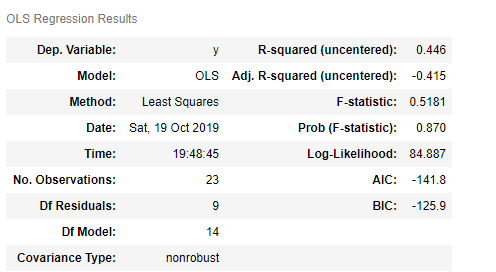
|  |  |  |
| --- | --- | --- |
| ***There is significant negative correlation between the target variables(PCT<>FWD) the CP rate and also the tbill-CP spread as can be seen above. This is further confirmed by the heat map, the pairwise scatter plots and the bee swarm plot that follows below. Consequently these are used as the features for further model buildings*** |  |  |
| **HeatMap** |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| Factor 12,13,14 – the target Variables  Factor 6-12 – the CP rate and its spreads with t-bills |  |  |
|  |  |  |
|  |  |  |
| **Pairwise Scatter Plots** |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Bee Swarm Plot**

Train/Test Split and feature Scaling

A 90/10 train test split is taken with random state = 42 and these parameters are kept constant across all the three target variables. Features are scaled before proceeding to individual models.

1. **Model Fitting: For three month Fwd- PCT3MOFWD**
2. **Linear regression model**



**The R2 is only about 45% . Lets move on to the next model – regression tree**

1. **Regression Tree Model**

**max\_depth=4,min\_samples\_leaf=0.1,random\_state=3**

**RMSE:** 0.00385

**The regression Tree model fits extremely well with a very low rmse**

**Using a 10 fold cross validation to improve the model reduces the MSE.**

Train MSE: 0.0000144

Test MSE: 0.00001485

1. **SVR**

R-square:-0.005509784364353454

**R2 is negative which suggests that this model is arbitrarily worse.**

**5 .Ensembling – Random Forrest regressor**

**n\_estimators=400,min\_samples\_leaf=0.12,random\_state=1**

Test set RMSE of rf: 0.004138

The Random Forest Regressor does a very good job in training the individual trees and introduces further randomization

**Similar analysis were carried out for the other two target variables -** PCT6MOFWD & PCT9MOFWD

**Summarizing the results:**

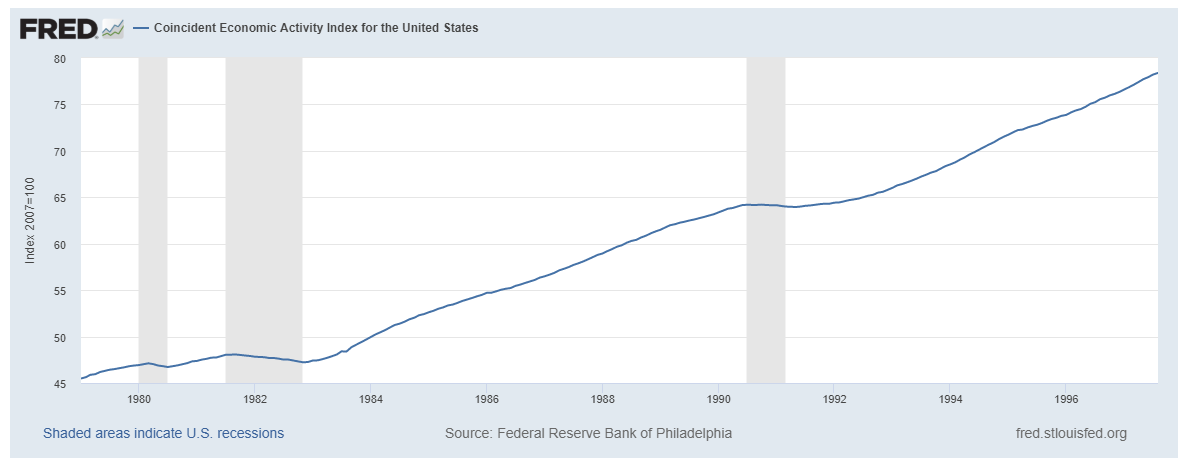
|  |  |  |  |
| --- | --- | --- | --- |
| Model Used | PCT3MOFWD | PCT6MOFWD | PCT9MOFWD |
| Linear Regression(R^2 ) | 0.45 | 0.43 | 0.9 |
| Regression tree(RMSE ) | 0.00385 | 0.00161 | 0.0099 |
| Support Vector Regreesion(R^2) | -0.0055 | -0.007 | -0.028 |
| Random Forrest Regressor(RMSE) | 0.004138 | 0.00225 | 0.009566 |

**6.Conclusions**

From the study above we find that random forrest is the best model for this problem set. Linear Regression does well for predicting the longer term target variable – **PCT9MOFWD** but is less than 50% for the other two.   
On the other hand Regression Tree and random Forrest(which both use decision tree as its estimator) performs remarkably well.

The negative R^2 on SVR suggests this model is arbitrarily worse and is not considered for further analysis.

**Economic Rationale:**



The Commercial Paper rate is the short term borrowing rate in the repo market. As predicted by our models (and indeed by the graph above from the Federal Reserve Bank) , the cp rate and its spread over tbills is a good indicator of upcoming economic downturn. We can see the change in index above decreased over all the major recession periods.