IE598 - Machine Learning in Finance - Group Project

Members :

1. Rakesh Reddy Mudhireddy NetID : 'rmudhi2'
2. Ankur Mukherjee NetID : 'ankurm3'
3. Jianwei Su NetID : 'jianwei5'

Index :

Click on the page number directly to go to corresponding section

Chapter 1 : Credit Score Problem

1. Introduction :

- In this problem, we have credit score data with 1700 observations of 26 financial and accounting metrics changes for a set of firms in several different industries.
- The Class label is the Moody's credit rating assigned to the firm in the following quarter. Certain ratings are considered Investment Grade (=1), other ratings are not (=0) and consequently may not be held in certain institutional portfolios (pension plans, etc.)
- It is a classification problem, goal being classifying Investment Grade and Moody's Score for a firm. Investment Grade being binary variable and Moody's Score being multiclass variable. So, two different set of models are built, one set for binary classification and another set for multiclass classification.

2. Exploratory Data Analysis :

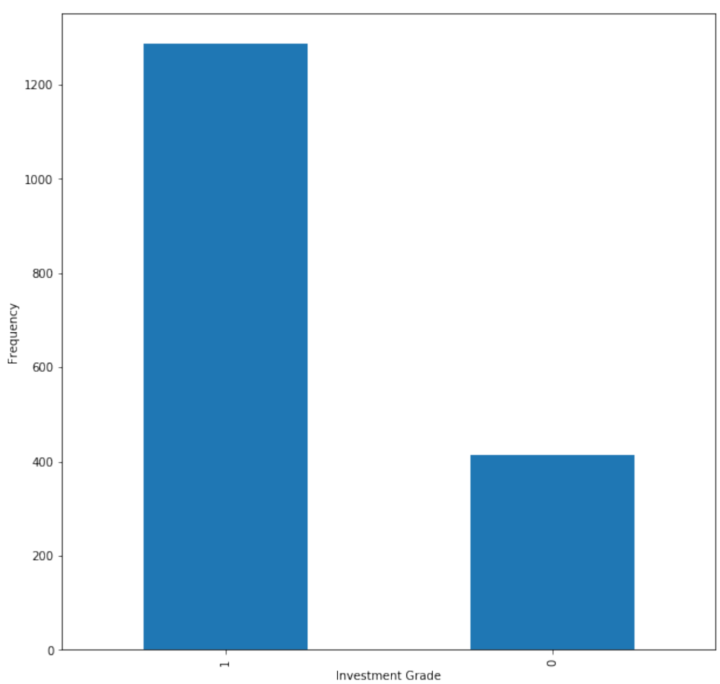
1. Summary Statistics :

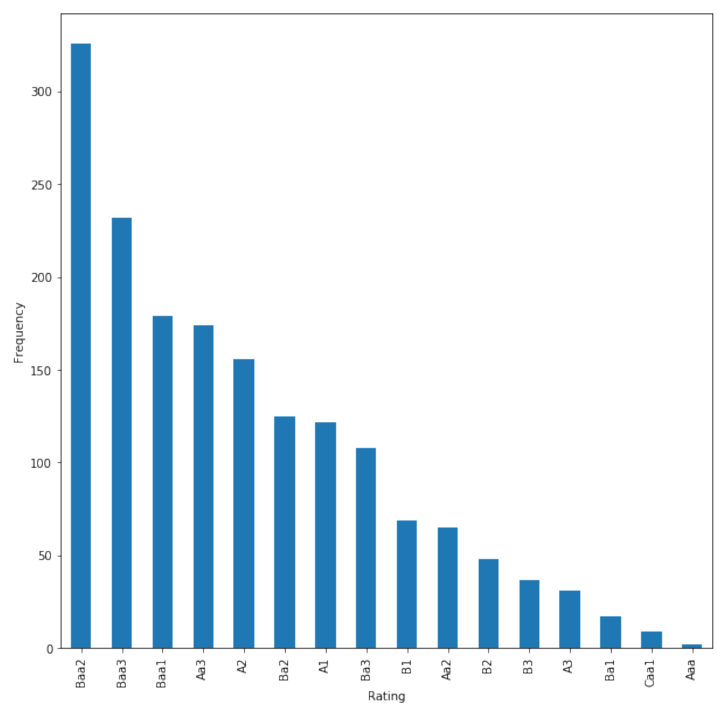
	Sales/Revenues	Gross Margin	EBITDA	EBITDA Margin	Net Income Before Extras	Total Debt	Net Debt	LT Debt	ST Debt	Cash	Free Cash Flow
count	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000
mean	0.050378	0.026007	0.068718	0.021074	0.123026	0.822405	-0.419810	1.255168	3.142797	0.466620	-0.312325
std	0.161910	0.273768	0.237365	0.189025	14.475689	13.317075	28.385702	16.224453	51.986550	1.859494	8.895136
min	-0.661715	-0.794722	-0.782254	-0.805153	-289.000000	-0.903014	-493.305578	-0.921515	-0.997692	-0.990982	-238.750000
25%	-0.005693	-0.020028	-0.022640	-0.042771	-0.158478	-0.076316	-0.120725	-0.094767	-0.337959	-0.195117	-0.527219
50%	0.034000	0.003403	0.049482	0.011134	0.056627	0.005886	-0.003060	-0.002078	0.043092	0.075820	-0.058475

Total Debt/EBITDA	Net Debt/EBITDA	Total MV	Total Debt/MV	Net Debt/MV	CFO/Debt	CFO	Interest Coverage	Total Liquidity	Current Liquidity	Current Liabilities
1700.000000	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000
0.731197	-0.819863	0.092043	1.270202	-0.398624	-0.165088	-0.189317	0.298785	-0.855714	0.436002	0.072802
12.280493	22.002550	0.385111	22.797054	41.235876	6.277606	5.668669	5.265291	22.926862	1.904282	0.266471
-0.910486	-495.355952	-0.871567	-0.939190	-781.502439	-172.654240	-161.609425	-0.991976	-502.000000	-0.994141	-0.684678
-0.134477	-0.181621	-0.113241	-0.206442	-0.267345	-0.211115	-0.115159	-0.096996	-0.857013	-0.227327	-0.072734
-0.012302	-0.034452	0.066836	-0.018464	-0.032055	0.012847	0.046983	0.043216	-0.229098	0.040446	0.041785
0.141443	0.163697	0.236566	0.242868	0.274710	0.251992	0.216432	0.177340	0.512778	0.416067	0.161215
256.050232	360.926171	3.961121	676.443064	1352.088710	15.821709	13.005788	182.131887	280.138728	34.372455	4.194381

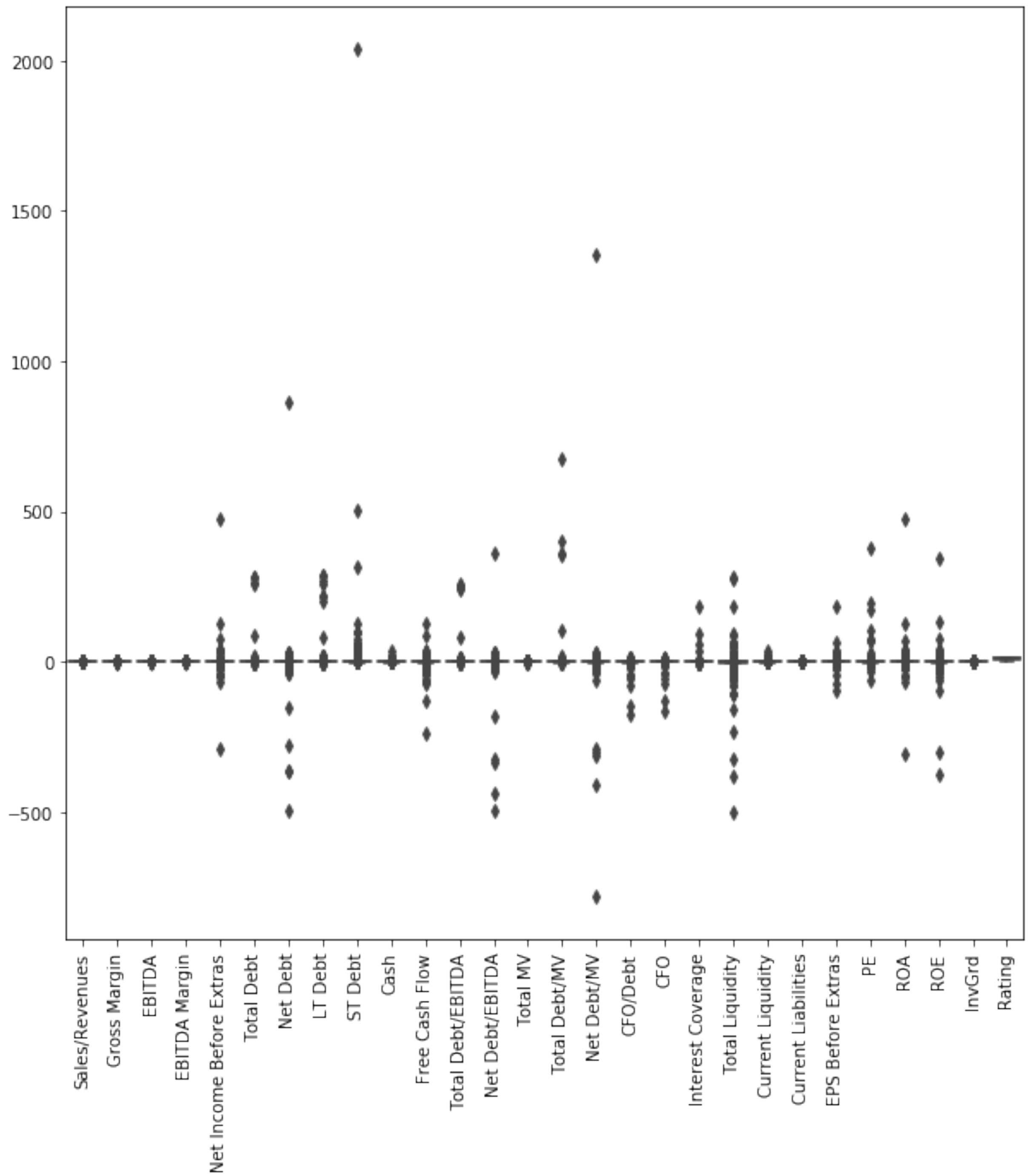
Current Liabilities	EPS Before Extras	PE	ROA	ROE	InvGrd
1700.000000	1700.000000	1700.000000	1700.000000	1700.000000	1700.000000
0.072802	0.032196	0.497705	0.019394	-0.217604	0.757059
0.266471	6.151994	12.102502	14.594193	15.389000	0.428986
-0.684678	-96.250000	-59.795133	-305.462167	-373.837267	0.000000
-0.072734	-0.152894	-0.293521	-0.208483	-0.233955	1.000000
0.041785	0.066027	-0.040405	-0.009403	-0.020392	1.000000
0.161215	0.236046	0.168897	0.156136	0.201596	1.000000
4.194381	187.000000	381.243282	474.847172	343.145356	1.000000

2. Histograms of Dependent variables :

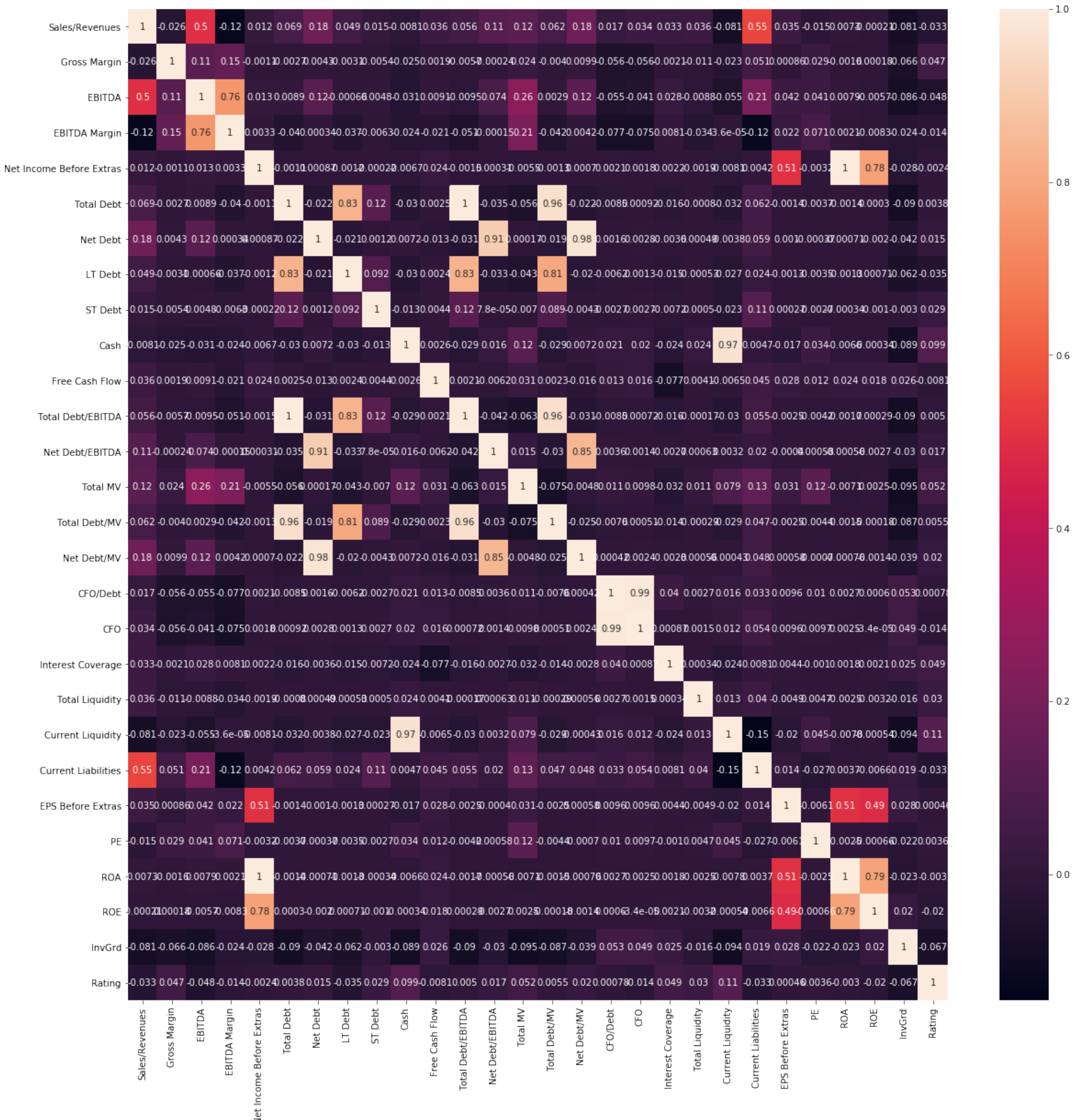




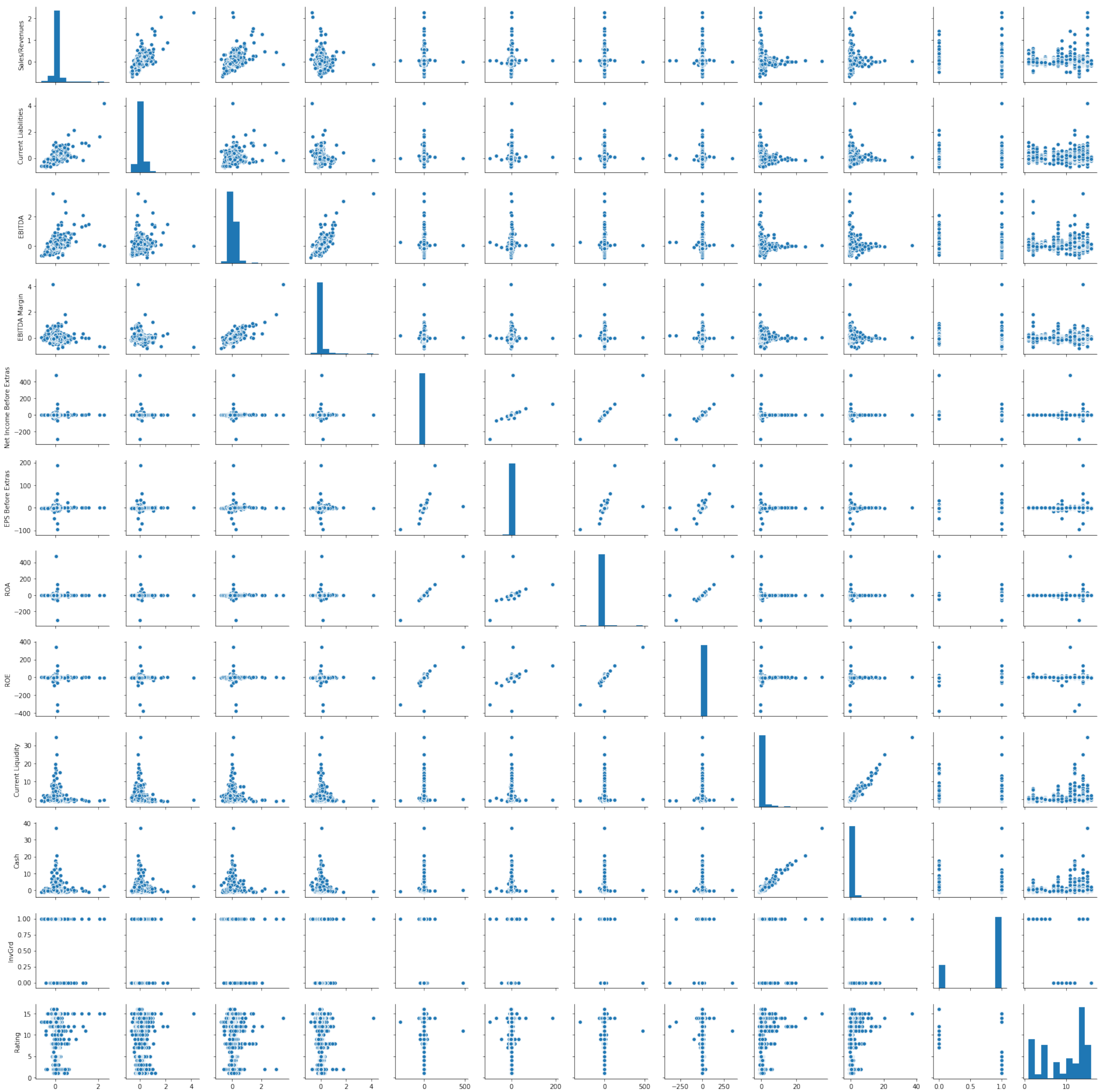
3. Box Plot :

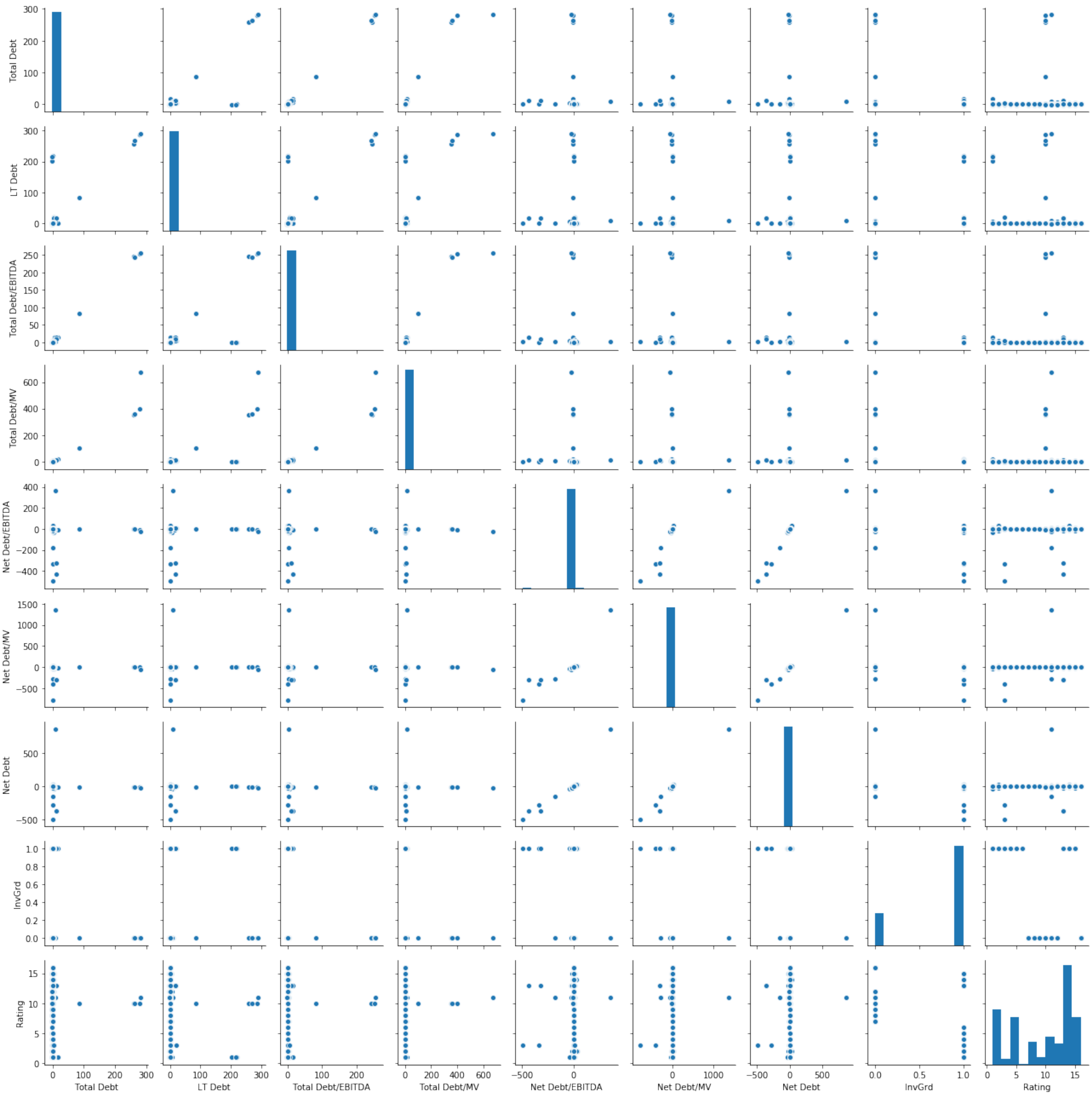

2000
1500
1000
500
0
−500
Sales/Revenues
Gross Margin
EBITDA
EBITDA Margin
Net Income Before Extras
Total Debt
Net Debt
LT Debt
ST Debt
Cash
Free Cash Flow
Total Debt/EBITDA
Net Debt/EBITDA
Total MV
Total Debt/MV
Net Debt/MV
CFO/Debt
CFO
Interest Coverage
Total Liquidity
Current Liquidity
Current Liabilities
EPS Before Extras
PE
ROA
ROE
InvGrd
Rating

4. Correlation Heat Map:



5. Scatter plots :




Total Debt
LT Debt
Total Debt/EBITDA
Total Debt/MV
Net Debt/EBITDA
Net Debt/MV
Net Debt
InvGrd
Rating

3. Preprocessing :

1. There are no missing values in dataset
2. Rating variable is categorical - so converted it into numerical values
3. Dataset splitting is done using random state and stratify (since there is imbalance in values of target variables)
4. Feature scaling is not essential for this dataset, as already values are comparable (one can verify this from summary statistics)

4. Feature Selection :

Feature selection is done using two methods:

1. *Correlation heat map and scatter plots*

- From scatter plots 1 - ('EBITDA', 'EBITDA Margin'), ('Net Income Before Extras', 'ROA', 'ROE'), ('Current Liquidity', 'Cash')
- These combinations are correlated, so out of these 7 variables, 3 variables can be selected :
- selection based on correlation with 'InvGrd' & 'Rating' :
- selected ones for 'InvGrd' : ('EBITDA', 'Net Income Before Extras', 'Current Liquidity')
- selected ones for 'Rating' : ('EBITDA', 'ROE', 'Current Liquidity')

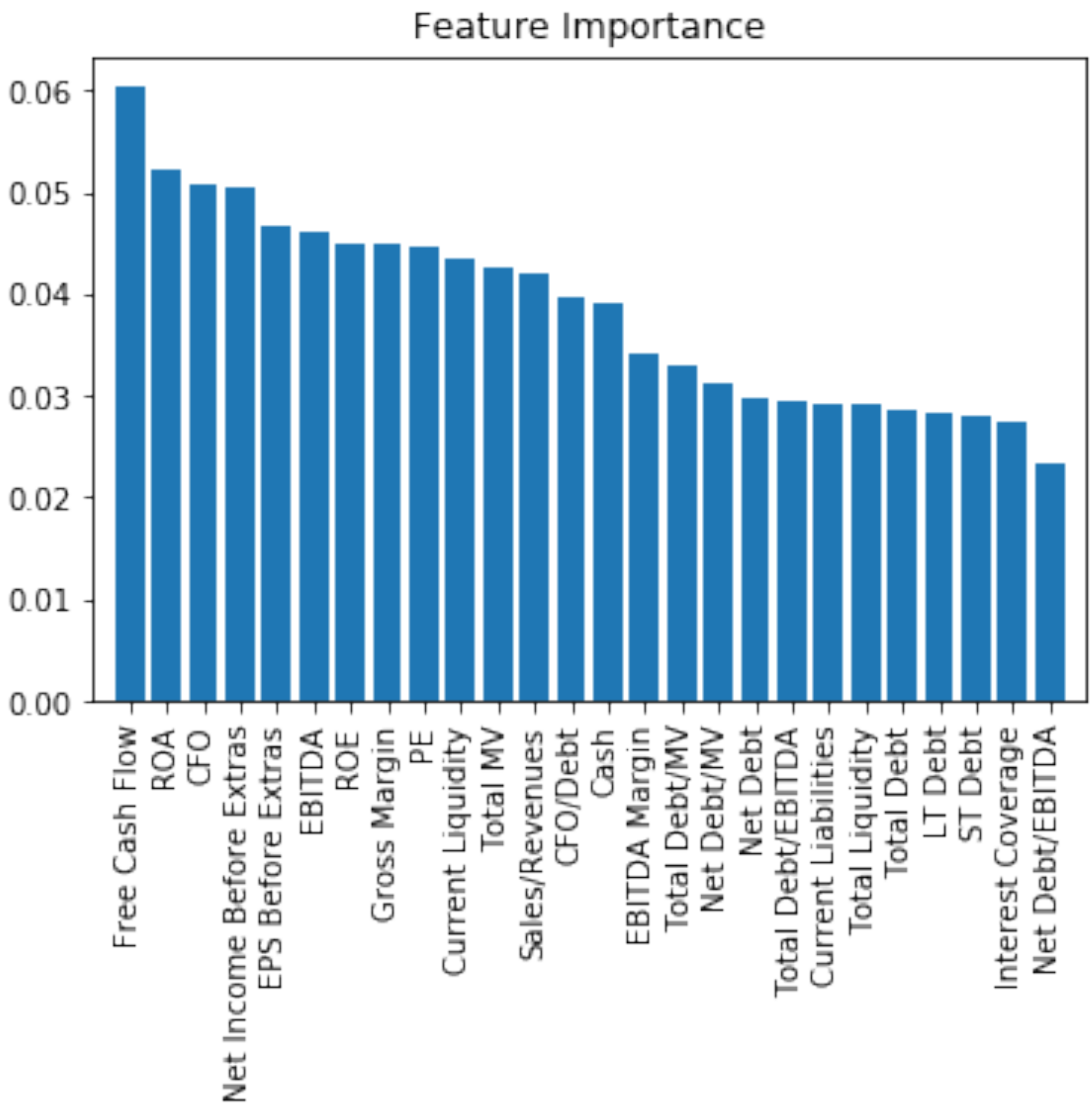
- From scatter plots 2 - ('Total Debt','LT Debt','Total Debt/EBITDA','Total Debt/MV'), ('Net Debt/EBITDA','Net Debt/MV', 'Net Debt')
- These combinations are correlated, so out of these 7 variables, 2 variables can be selected :
- selected ones for 'InvGrd' : ('Total Debt','Net Debt')
- selected ones for 'Rating' : ('Total Debt/MV','Net Debt/MV')

- So from these 14 variables, we can select 5 variables for our model, with minimal loss of explained variance

- Modeling has been done with these subsets of features as well, but there is no significant difference in model performance or computational time (dataset is small)

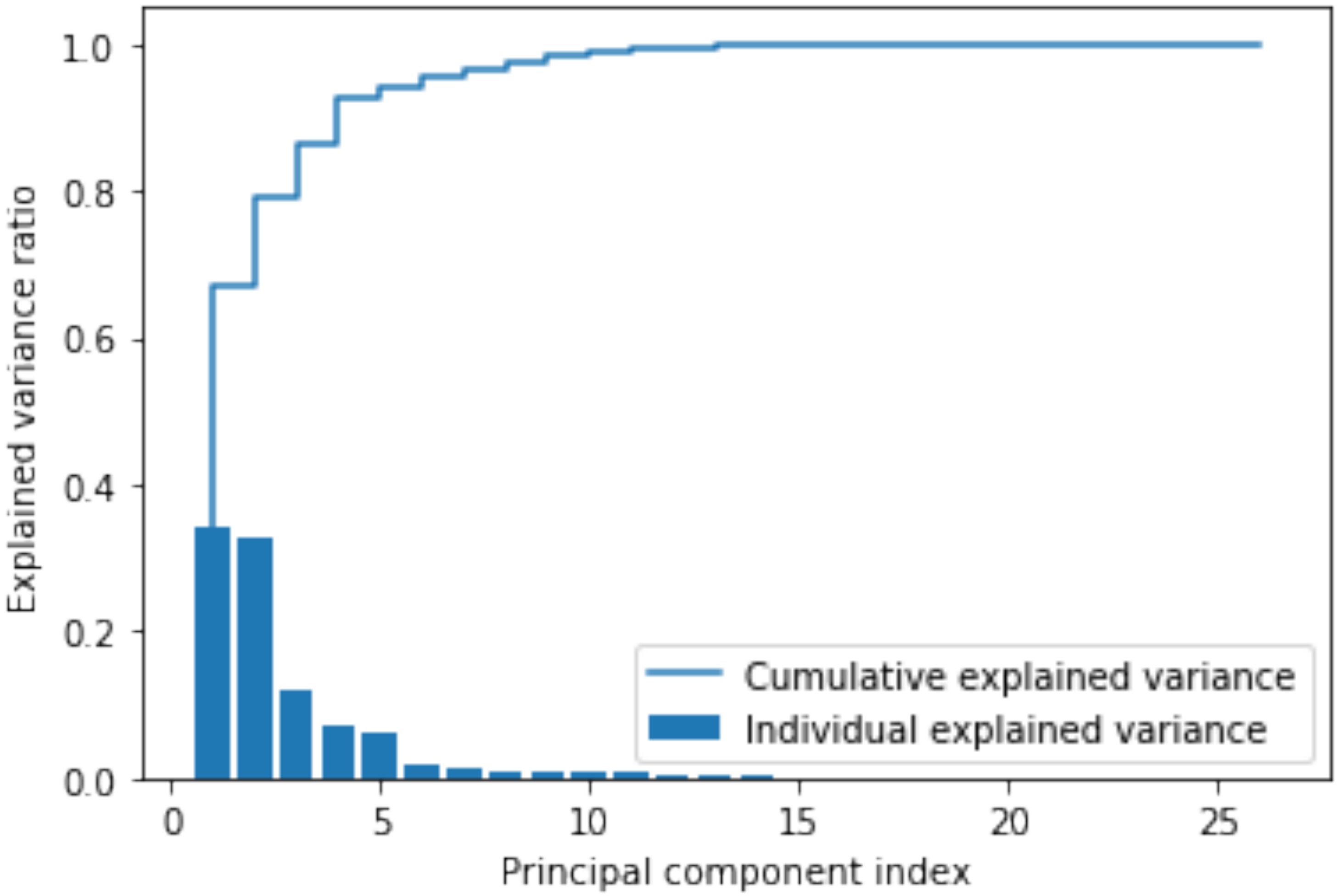
2. *Feature Importance using Random Forest Classifier :*

- Even the subset of features using this method, doesn't give a significantly different model



5. Feature Extraction :

- Feature extraction is done using Principal Component Analysis. These PCA features also do not give any significant difference in model performance.



6. Model fitting and evaluation/Hyperparameter tuning/Ensembling :

Binary Classification (Investment Grade)				
Model	***Best Hyperparameters***	***Test Accuracy***	***F1-score***	***Time taken (for GridSearchCV)***
Logistic Regression	Penalty = 'l2', C = 100	74.70%	0.852	2.19s
Logistic Regression with PCA	Penalty = 'l1', C = 1	74.70%	0.853	0.54s
KNN	n_neighbors = 5, p = 1, weights = 'distance'	78.82%	0.867	5.92s
KNN with PCA	n_neighbors = 5, p = 1, weights = 'distance'	78.82%	0.872	4.26s
Decision Tree Classifier (DTC)	Criterion = 'entropy', max_depth = 20	79.41%	0.859	2.85s
DTC with PCA	Criterion = 'entropy', max_depth = 10	76.47%	0.842	0.57s
SVM Classifier (SVC)	gamma = 'auto'	78.82%	0.876	0.53s
Random Forest Classifier (Ensembling)	criterion = 'entropy', max_depth = 20, n_estimators = 500	85.88%	0.91	155.85s

Multiclass Classification (Rating)			
Model	***Best Hyperparameters***	***Test Accuracy***	***Time taken (for GridSearchCV)***
Logistic Regression	Penalty = ‘l1’, C = 100	21.76%	49.56s
Decision Tree Classifier (DTC)	Criterion = ‘entropy’, max_depth = 20	47.05%	2.81s
KNN	n_neighbors = 5, p = 1, weights = ‘distance’	47.05%	5.45s
SVM Classifier (LinearSVC)	None given	20.58%	N/A
Random Forest Classifier (Ensembling)	criterion = ‘gini’, max_depth = 50, n_estimators = 300	70.58%	298.05s

7. Conclusions :

- Logistic Regression, KNN, Decision Trees & SVM Classifier are used in binary classification case and for ensembling Random Forest classifier is used.
- F1-score is considered as evaluation metric along with accuracy score, as the investment grade data is biased towards value ‘1’.
- Based on f1-score and accuracy score on test data, Random forest classifier gives best performance for this dataset.
- In the case of multiclass classification, Logistic Regression, Decision Trees, KNN, LinearSVC & for ensembling Random Forest Classifier are used to build models.
- Even in this case, Random Forest classifier performs well with this dataset.

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 Validation
- 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 modeled and fitted with the attributes. The results are studied for each and an inference derived.

2. Exploratory Data Analysis & Preprocessing:

1. Summary Statistics :

The following table shows a basic summary of the dataset:

	T1Y Index	T2Y Index	T3Y Index	T5Y Index	T7Y Index	T10Y Index	CP1M	CP3M	CP6M	CP1M_T1Y	CP3M_T1Y	CP6M_T1Y	PCT3MOFWD	PCT6MOFWD	PCT9MOFWD
0	10.41	9.86	9.50	9.20	9.14	9.10	9.75	9.95	10.01	0.936599	0.955812	0.961575	0.011470	0.018060	0.024406
1	10.24	9.72	9.29	9.13	9.11	9.10	9.74	9.90	9.96	0.951172	0.966797	0.972656	0.009298	0.014866	0.020612
2	10.25	9.79	9.38	9.20	9.15	9.12	9.72	9.85	9.87	0.948293	0.960976	0.962927	0.010340	0.015455	0.020154
3	10.12	9.78	9.43	9.25	9.21	9.18	9.86	9.95	9.98	0.974308	0.983202	0.986166	0.006720	0.013141	0.017409
4	10.12	9.78	9.42	9.24	9.23	9.25	9.77	9.76	9.71	0.965415	0.964427	0.959486	0.005653	0.011451	0.016353

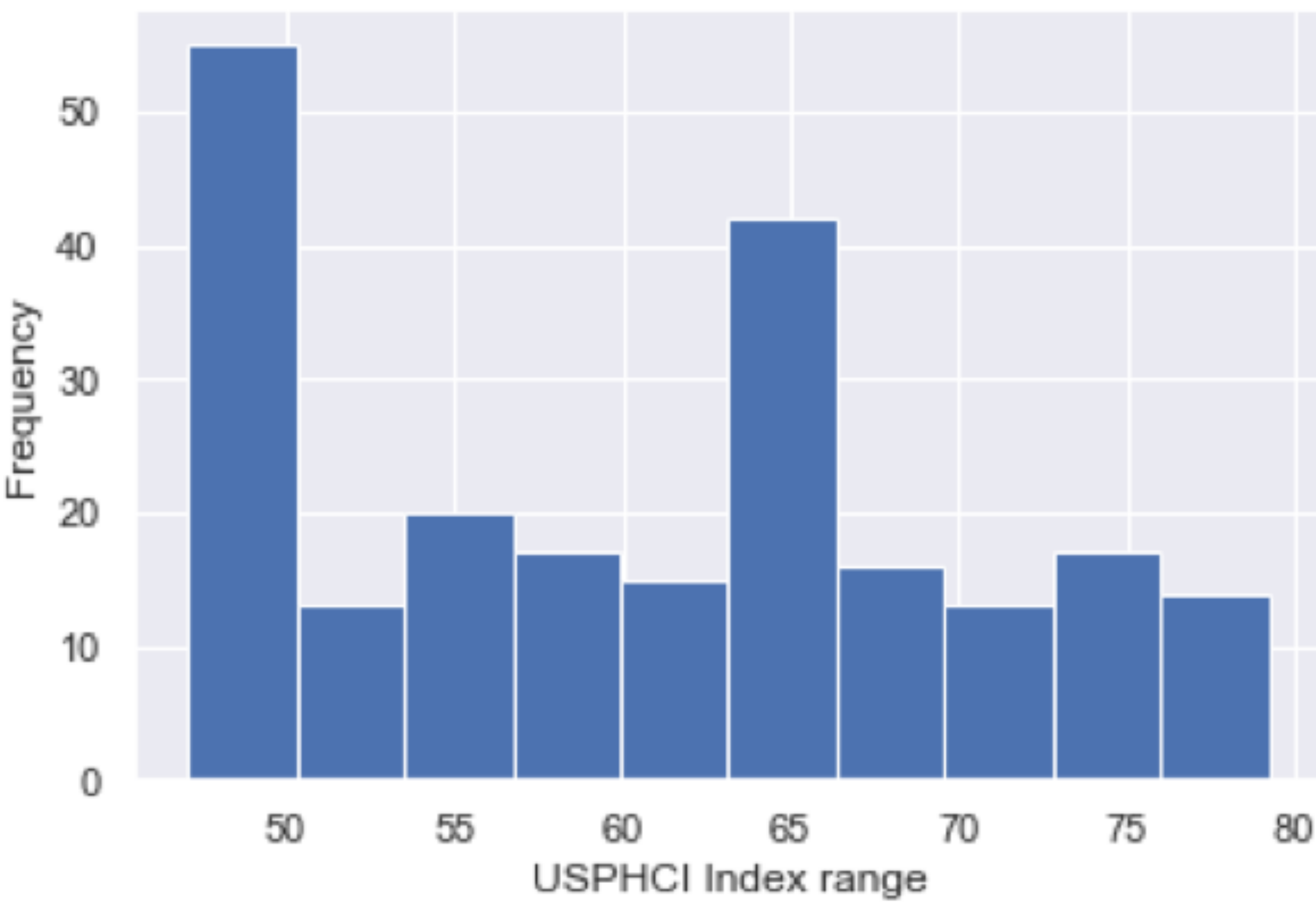
The basic characteristics of the features was found out using the describe function. As an example:

```
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.

2. Histogram for the USHPCI Index:

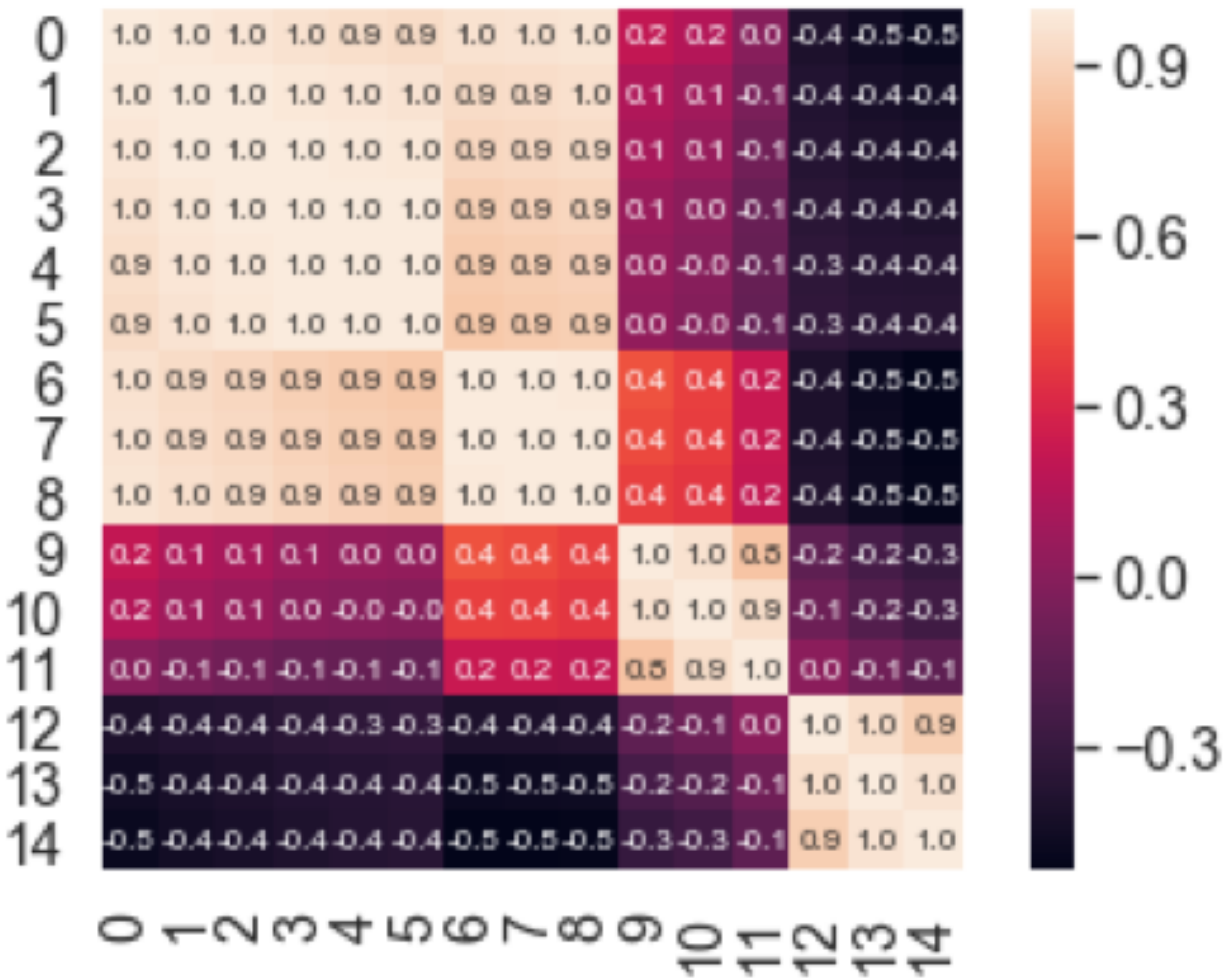


3. Correlations:

	T1Y Index	T2Y I	CP1M	CP3M	CP6M	CP1M_T1Y	CP3M_T1Y	CP6M_T1Y
T1Y Index	1.000000	0.99	0.962641	0.967578	0.972892	0.208129	0.152748	0.001319
T2Y Index	0.992364	1.00	0.938317	0.945106	0.954110	0.142681	0.089627	-0.050497
T3Y Index	0.981588	0.99	0.920249	0.927676	0.938247	0.109464	0.057786	-0.075841
T5Y Index	0.962056	0.98	0.891523	0.899770	0.912096	0.063193	0.013663	-0.111213
T7Y Index	0.947012	0.97	0.873208	0.881932	0.895172	0.045970	-0.001902	-0.122058
T10Y Index	0.935681	0.96	0.860518	0.869407	0.883008	0.034947	-0.011444	-0.127925
CP1M	0.962641	0.93	1.000000	0.998395	0.993283	0.449292	0.393453	0.229515
CP3M	0.967578	0.94	0.998395	1.000000	0.997943	0.427221	0.383779	0.231517
CP6M	0.972892	0.95	0.993283	0.997943	1.000000	0.393722	0.358950	0.221016
CP1M_T1Y	0.208129	0.14	0.449292	0.427221	0.393722	1.000000	0.960717	0.842279
CP3M_T1Y	0.152748	0.08	0.393453	0.383779	0.358950	0.960717	1.000000	0.946781
CP6M_T1Y	0.001319	-0.05	0.229515	0.231517	0.221016	0.842279	0.946781	1.000000
PCT3MOFWD	-0.406827	-0.38	-0.404316	-0.401550	-0.395089	-0.150104	-0.096440	0.010641
PCT6MOFWD	-0.454663	-0.42	-0.475103	-0.471441	-0.463500	-0.239304	-0.192718	-0.083744
PCT9MOFWD	-0.483731	-0.44	-0.520132	-0.515032	-0.505840	-0.303912	-0.259039	-0.149062

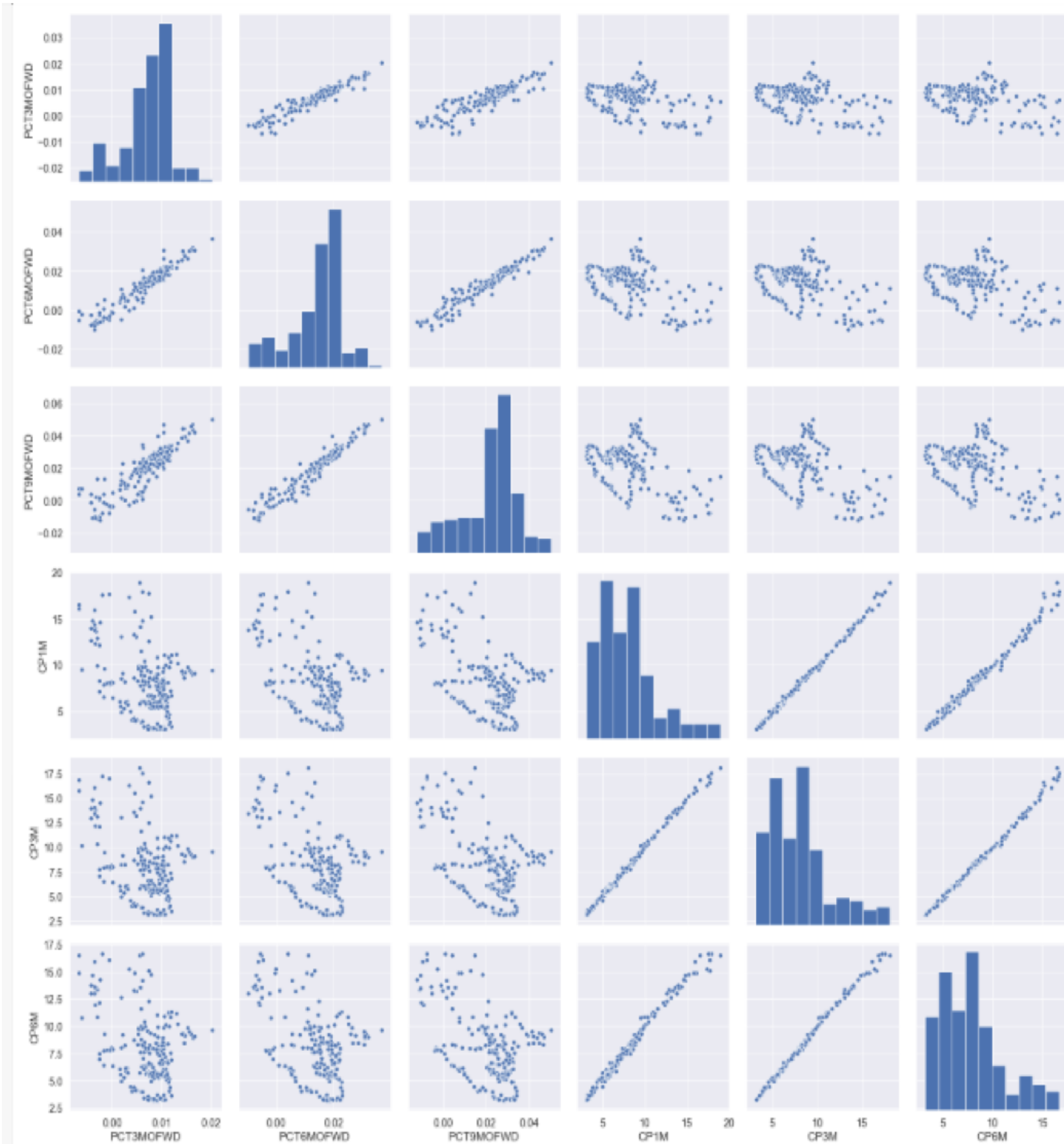
There is significant negative correlation between the target variables(PCT<>FWD) the CP rate and also the tbill-CP spread as can be seen above. This is further confirmed by the heat map, the pairwise scatter plots and the bee swarm plot that follows below. Consequently these are used as the features for further model buildings

4. Correlation Heatmap :

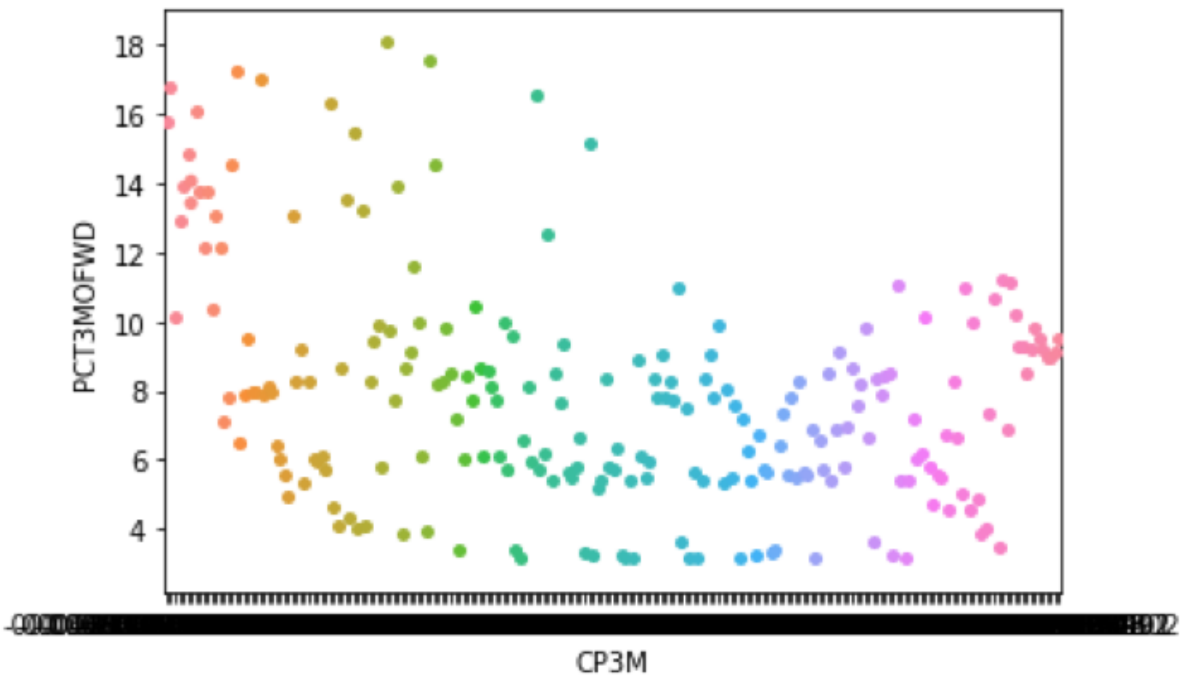


Factor 12,13,14 – the target Variables
Factor 6-12 – the CP rate and its spreads with t-bills

5. Pairwise Scatter Plots:



6. 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.

3. Model Fitting: For three month Fwd-PCT3MOFWD:

I. Linear regression model

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.446
Model:	OLS	Adj. R-squared (uncentered):	-0.415
Method:	Least Squares	F-statistic:	0.5181
Date:	Sat, 19 Oct 2019	Prob (F-statistic):	0.870
Time:	19:48:45	Log-Likelihood:	84.887
No. Observations:	23	AIC:	-141.8
Df Residuals:	9	BIC:	-125.9
Df Model:	14		
Covariance Type:	nonrobust		

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

II. 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

III. SVR

R-square:-0.0055097843643535454

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

4. 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

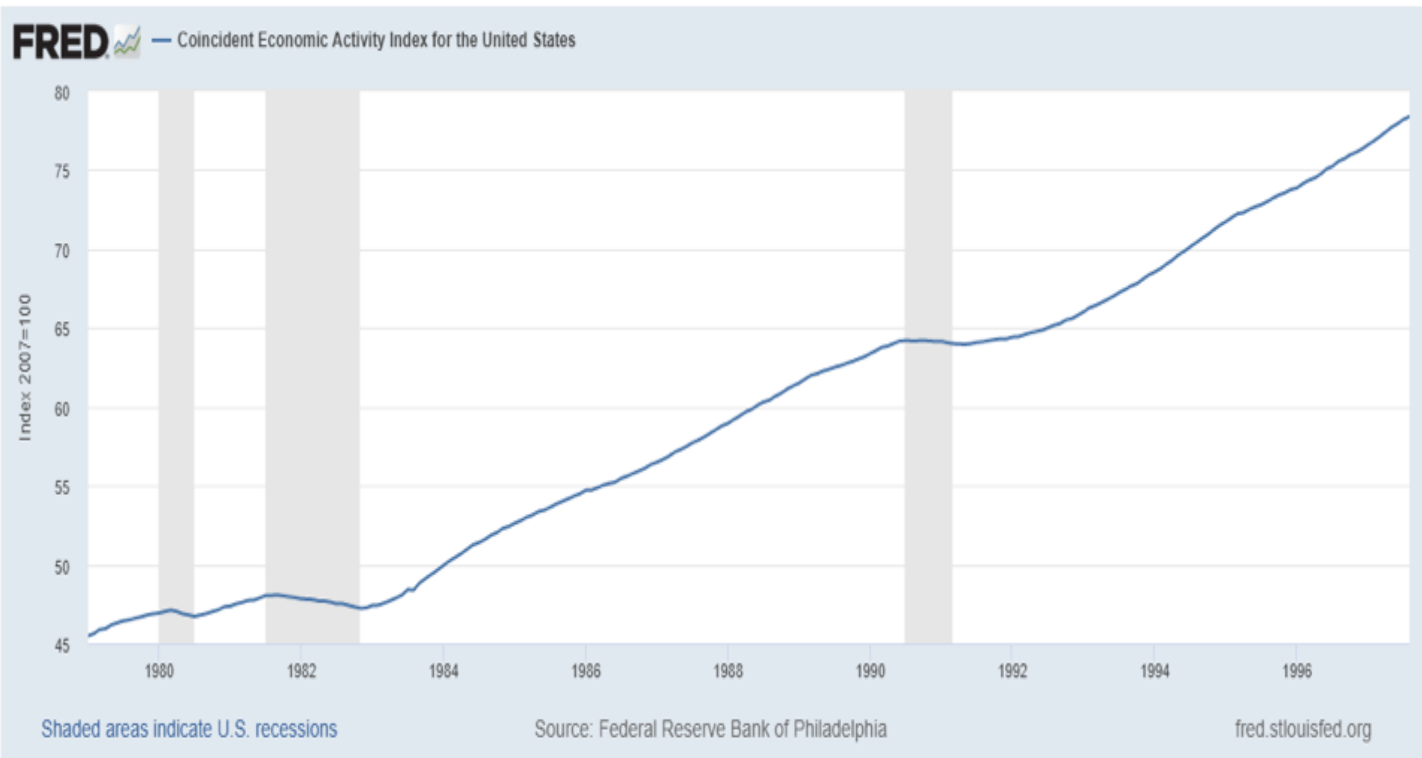
Summarizing the results:

Model	***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

5. 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 t-bills is a good indicator of upcoming economic downturn. We can see the change in index above decreased over all the major recession periods.

Appendix :

Chapter 1 GitHub link :

https://github.com/rakesh1827/IE598MLF_Group_project/blob/master/MLF_GP1_CreditScore.py

Chapter 2 GitHub link :

https://github.com/rakesh1827/IE598MLF_Group_project/blob/master/MLF_GP2_EconCycle.py