IE598 - MLF Final Project

Fall 2018

by Yuchen Duan, Ruozhong Yang, Fengkai Xu, Biao Feng, and Joseph Loss

**10/21/2018**

**University of Illinois at Urbana-Champaign**

**IE598 - Machine Learning in Finance**

Contents

[Chapter 1: Moody’s Bond Rating Classifier 1](#_Toc527931517)

[Exploratory Data Analysis 1](#_Toc527931518)

[Preprocessing & Feature Extraction/Selection 2](#_Toc527931519)

[Model Fitting & Evaluation(Binary&muticlasses) 3](#_Toc527931520)

[Hyperparameter Tuning 3](#_Toc527931521)

[Ensembling 4](#_Toc527931522)

[Conclusions 4](#_Toc527931523)

[Chapter 2: USPHCI Economic Activity Forecast 5](#_Toc527931524)

[Exploratory Data Analysis 5](#_Toc527931525)

[Preprocessing & Feature Extraction/Selection 6](#_Toc527931526)

[Model Fitting & Evaluation 8](#_Toc527931527)

[Hyperparameter Tuning 8](#_Toc527931528)

[Ensembling 11](#_Toc527931529)

[Conclusions 11](#_Toc527931530)

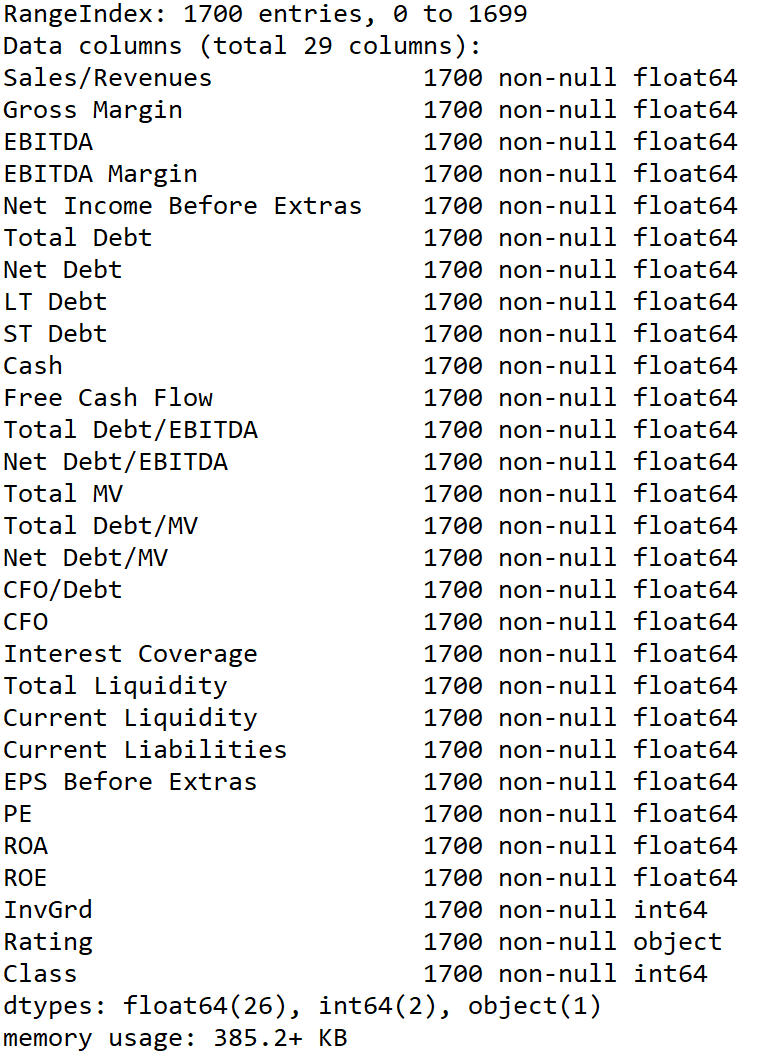
[Appendix 12](#_Toc527931531)

[Github Repository 12](#_Toc527931532)

# Chapter 1: Moody’s Bond Rating Classifier

## Exploratory Data Analysis

Here is what our data looks like:



Also, we have a correlation matrix:

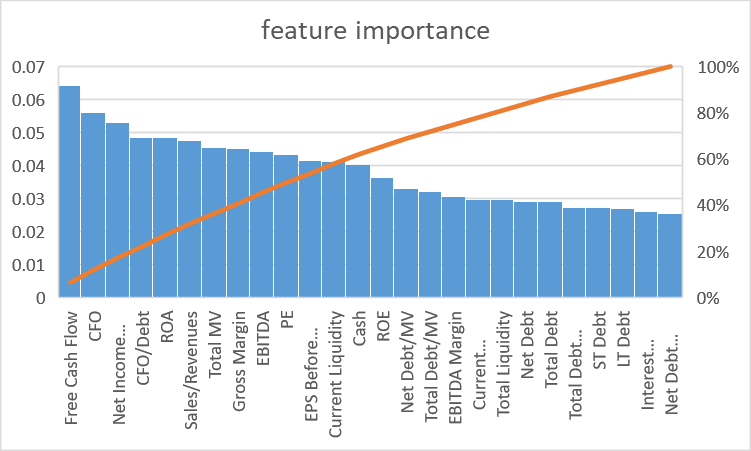


## Preprocessing & Feature Extraction/Selection

The preprocessing part combine some steps that need to be done before we try to fit our model:

1. Split the test and train database via train\_test\_split(with test\_size = 0.1 and random\_state=42)
2. Standardize features via StandardScaler for better model performance.

We also calculate the importance of each feature and select 13 of them for our models.



## Model Fitting & Evaluation(Binary&muticlasses)

1. Model 1

The first model is the KNN model.

1. Model 2

The second model is the Random Forest model.

1. Model 3

The third model is the Decision Tree model.

1. Model 4

The forth model is the Logistic Regression model.

We will discuss those models in the hyperparameter tuning and ensemble parts.

## Hyperparameter Tuning

We deal with different parameters via GridSearchCV function, the range of each model’s parameter is form 1 to 100. Here is the best result for each model:



From the table, it is easy to find the muitclasses task lead to poor prediction(muti\_lr score is about 1/3 compare to the binary one). There are several improvements can be done for better models, we will discuss them at the conclusion.

## Ensembling

Our team use the ensembling method for binary classification does not support muticlassification task. Result showed below:



## Conclusions

The best result for binary model is 0.89(after ensembling) and the best for multiclass is 0.67. There are several things we can do to improve our model:

Dimension reduction

We can reduce the dimension of our model for better prediction, but may let miss some important information.

Internal relationships

Some features are highly correlated, we can find them and just use one of them. Besides, many features have internal relationships, thus, some of them may actually talk about the same thing.

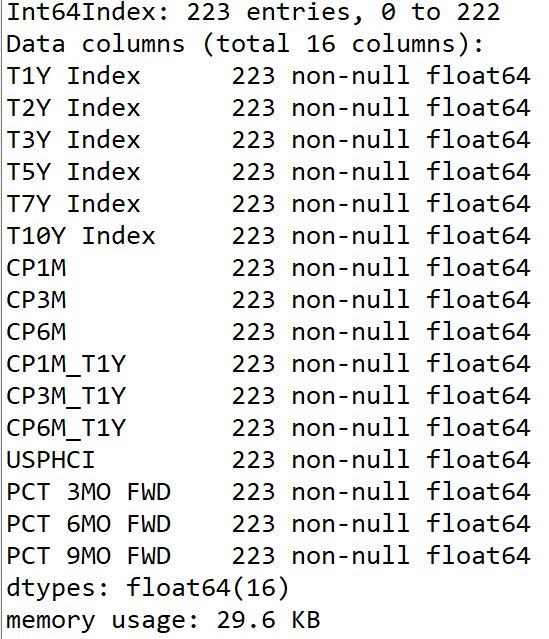
Weight adjustment

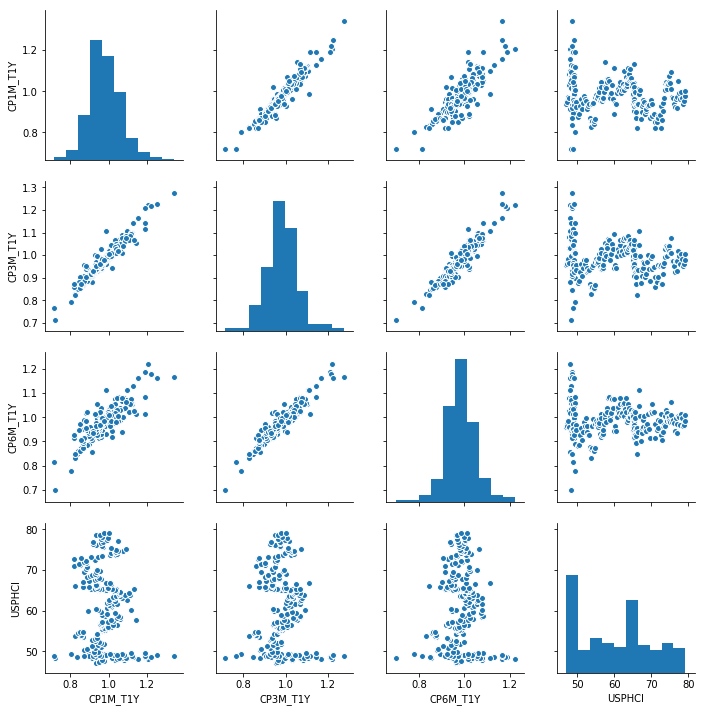
Although those models adjust weights of each feature automatically, people from accounting major may hold different view of those weights.

# Chapter 2: USPHCI Economic Activity Forecast

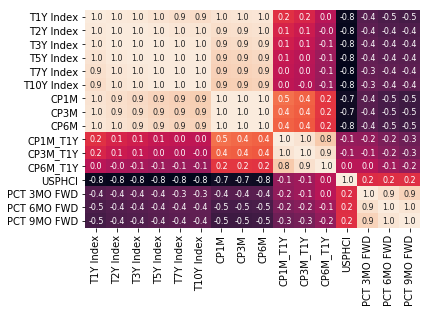
## Exploratory Data Analysis

Our data looks like this:



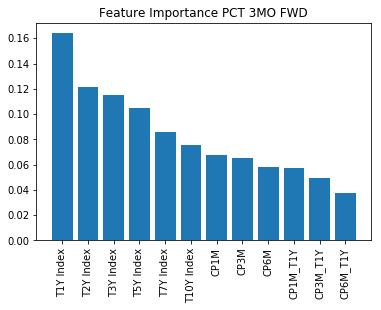


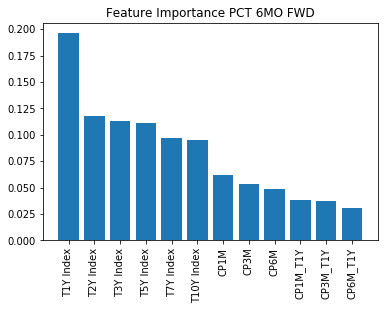
Also, we need to see the relation between each features:

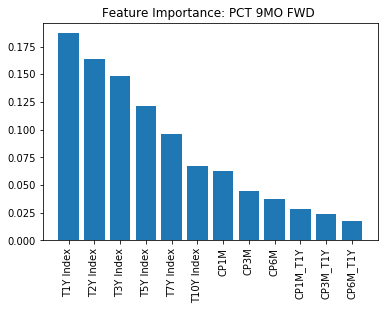


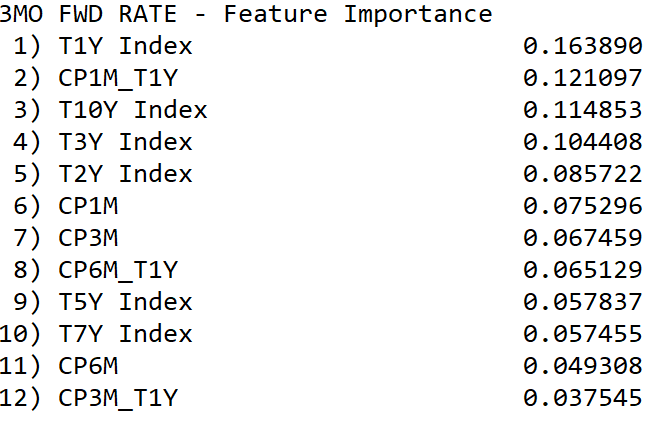
## Preprocessing & Feature Extraction/Selection

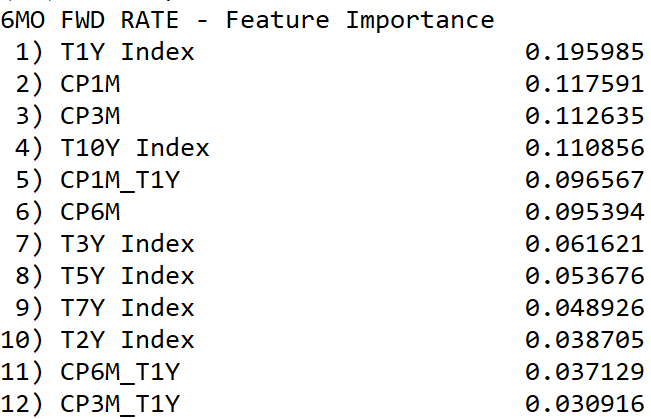
We can see the importance of each feature in all three situations:













## Model Fitting & Evaluation

1. Model 1

We use Linear Regression for 3-month situation.

1. Model 2

We use Ridge Regression for 6-month situation.

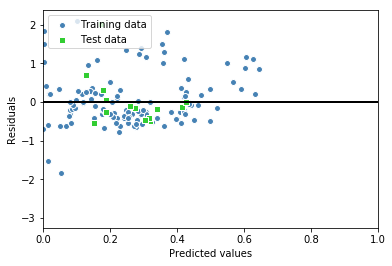
1. Model 3

We use Lasso Regression for 9-month situation.

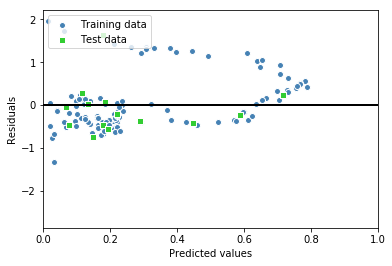
## Hyperparameter Tuning

In the first case(linear regression), we cannot change the parameter, in the second and third cases, we change the alpha(ridge from 10^-3 to 10^0, lasso from 10^-6 to 10^-3).We only show those images for the best model of each case and show the rest of them in a table.

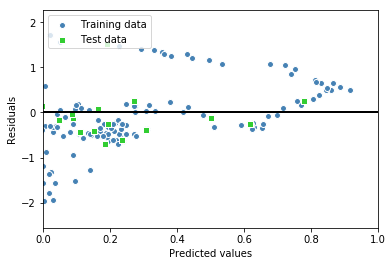
Linear Regression:



Ridge Regression:( Ridgealpha: 0.010)



Lasso Regression:( Lassoalpha: 0.000100)



Here is the table:



## Ensembling

After the ensembling process, we get better result:

Test set MSE: 0.31

Test set R-Squared: 0.64

## Conclusions

INSERT TEXT HERE

# Appendix

## Github Repository

[IE598\_F18\_MLF\_GROUP\_PROJECT (LINK)](https://github.com/chicago-joe/IE598_F18_MLF_GP/tree/master)