FE-520

FINAL PROJECT REPORT

CREDIT RATING PREDICTION USING MACHINE LEARNING

Submitted by Team 4–

Pinakini Samant

Sushant Ghorpade

Table of Contents

Figures 2

Introduction 3

DATA UNDERSTANDING 4

1. About Data 4
2. Variables/features 4
3. Data Set 5
4. Data Cleaning 5

MODELLING 7

SUPPORT VECTOR MACHINE 9

RANDOM FOREST 10

K Neighbours Classification 11

PRINCIPAL COMPONENT ANALYSIS 12

Performance of the Models on the Dataset 13

Conclusion and Future Scope 16

Bibliography 17

Figures

Figure1. Overview of dataset 05

Figure2. Data Structure 06

Figure3. Machine Learning Modelling process 07

Figure4. Credit Rating 08

Figure5. Support vector machine 09

Figure6. Random Forest 10

Figure7. Decision tree 10

Figure8. K-Nearest Neighbors 11

Figure9. Principal Component Analysis 12

Figure10. Dimensional reduction 14

Introduction

Credit ratings are used by investors, intermediaries such as investment banks, issuers of debt, and businesses and corporations. Both institutional and individual investors use credit ratings to assess the risk related to investing in a specific issuance, ideally in the context of their entire portfolio. Intermediaries such as investment bankers utilize credit ratings to evaluate credit risk and further derive pricing of debt issues. Debt issuers such as corporations, governments, municipalities, etc., use credit ratings as an independent evaluation of their creditworthiness and credit risk associated with their debt issuance. The ratings can, to some extent, provide prospective investors with an idea of the quality of the instrument and what kind of interest rate they should be expecting from it. Businesses and corporations that are looking to evaluate the risk involved with a certain counterparty transaction also use credit ratings. They can help entities that are looking to participate in partnerships or ventures with other businesses evaluate the viability of the proposition. (1)

A credit rating is an opinion of a particular credit agency regarding the ability and willingness an entity (government, business, or individual) to fulfill its financial obligations in completeness and within the established due dates. A credit rating also signifies the likelihood a debtor will default. It is also representative of the credit risk carried by a debt instrument – whether a loan or a bond issuance. (1)

The purpose of this is to classify customers in two categories i.e. good credit rating and bad credit rating. It is crucial since it deals with a lot of money into financial markets. To predict the credit rating with accuracy is very profitable to any finance market. Due to the growing competition to create a model with good accuracy has been a crucial part of any finance company.

DATA UNDERSTANDING

1. About Data

The data used for the project can be divided into two parts. The Historical Credit Ratings which were obtained from the Thomson Reuters Eikon Database. And the Historical Fundamental Data was obtained from the Data Understanding. Dataset can be found at https://simfin.com

The dataset consists of 1430 ratings of 400 companies considered over a period of 6 years from 2010 to 2016. The data for each company consists of quarterly values of 49 financial quantities over the specified period from 2010 to 2016.

2. Variables/features

Below is the list of variables present in our dataset. To note that some variables have " Curr" and "prev" in the end. "Curr" denotes current while "prev" denotes previous.

1. Total Equity curr
2. Current Liabilities prev
3. Total Equity prev
4. Preferred Equity curr
5. Retained Earnings curr
6. Minorities curr
7. Equity Before Minorities curr
8. Long Term Debt curr
9. Equity Before Minorities prev
10. Intangible Asset curr
11. Retained earnings prev
12. Cash and Cash equivalent prev
13. Dividends curr
14. Accounts Payable curr
15. Debt to Assets Ratio prev
16. Net Change in PP&E & Intangibles prev
17. Debt to Assets Ratio curr
18. Operating Margin prev

19. Total Assets curr

20. Return on Assets prev

3. Data Set

```
data
```

7]:

	Revenues_prev	COGS_prev	SG&A_prev	R&D_prev	EBIT_prev	EBITDA_prev	Interest expense, net_prev	Abnormal Gains/Losses_prev	Income Taxes_prev	Net Income from Discontinued Op._prev	...	
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
7	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	14
8	4595.000000	1127.000000	1247.25	694.500000	1526.250000	1813.750000	21.000000	72.00	112.500000	0.00	...	14
9	4697.500000	1145.250000	1338.00	713.750000	1500.500000	1724.750000	69.500000	84.50	301.000000	0.00	...	8
10	4990.000000	1106.500000	1931.00	824.250000	1128.250000	1324.750000	97.750000	88.00	148.750000	0.00	...	17
11	5714.750000	1125.000000	1596.75	1071.250000	1921.750000	2130.750000	171.500000	37.50	375.250000	0.00	...	16
12	15.000000	11.000000	0.00	0.000000	2.000000	NaN	0.000000	0.00	0.000000	0.00	...	
13	53.500000	44.250000	0.00	0.000000	3.000000	NaN	1.750000	10.25	-2.000000	0.00	...	
14	101.500000	71.250000	0.00	0.000000	17.250000	NaN	7.000000	1.50	2.500000	-0.25	...	
15	177.750000	123.000000	0.00	0.000000	30.250000	33.333333	8.750000	3.50	6.000000	0.00	...	
16	250.500000	171.500000	0.00	0.000000	42.750000	NaN	11.500000	-1.00	10.000000	-0.25	...	
17	448.000000	300.500000	0.00	0.000000	79.750000	95.000000	26.250000	11.75	12.750000	-1.00	...	
18	891.250000	499.250000	188.50	73.500000	83.500000	153.500000	20.000000	7.50	16.000000	0.00	...	1

4. Data Cleaning

Handling of missing values - Due to unavailability of complete fundamental data (NAN values), the dataset was trimmed to contain 1198 rows.

```
Current Assets_curr                          1129 non-null float64
Net PP&E_curr                                1204 non-null float64
Intangible Assets_curr                       1179 non-null float64
Goodwill_curr                                1179 non-null float64
Total Noncurrent Assets_curr                 1129 non-null float64
Total Assets_curr                            1204 non-null float64
Short term debt_curr                         1204 non-null float64
Accounts Payable_curr                        1129 non-null float64
Current Liabilities_curr                     1129 non-null float64
Long Term Debt_curr                          1204 non-null float64
Total Noncurrent Liabilities_curr            1129 non-null float64
Total Liabilities_curr                       1204 non-null float64
Preferred Equity_curr                        1204 non-null float64
Share Capital_curr                           1204 non-null float64
Treasury Stock_curr                          1204 non-null float64
Retained Earnings_curr                       1204 non-null float64
Equity Before Minorities_curr                1204 non-null float64
Minorities_curr                              1204 non-null float64
Total Equity_curr                            1204 non-null float64
Depreciation & Amortisation_curr             1184 non-null float64
Change in Working Capital_curr               1184 non-null float64
Cash From Operating Activities_curr          1184 non-null float64
Net Change in PP&E & Intangibles_curr        1184 non-null float64
Cash From Investing Activities_curr          1184 non-null float64
Cash From Financing Activities_curr          1184 non-null float64
Net Change in Cash_curr                      1184 non-null float64
Free Cash Flow_curr                          1125 non-null float64
Gross Margin_curr                            1078 non-null float64
Operating Margin_curr                        1153 non-null float64
Net Profit Margin_curr                       1153 non-null float64
Return on Equity_curr                        1141 non-null float64
Return on Assets_curr                        1141 non-null float64
Current Ratio_curr                           1129 non-null float64
Liabilities to Equity Ratio_curr             1204 non-null float64
Debt to Assets Ratio_curr                    1204 non-null float64
rating                                       1222 non-null object
dtypes: float64(98), object(1)
```

MODELLING

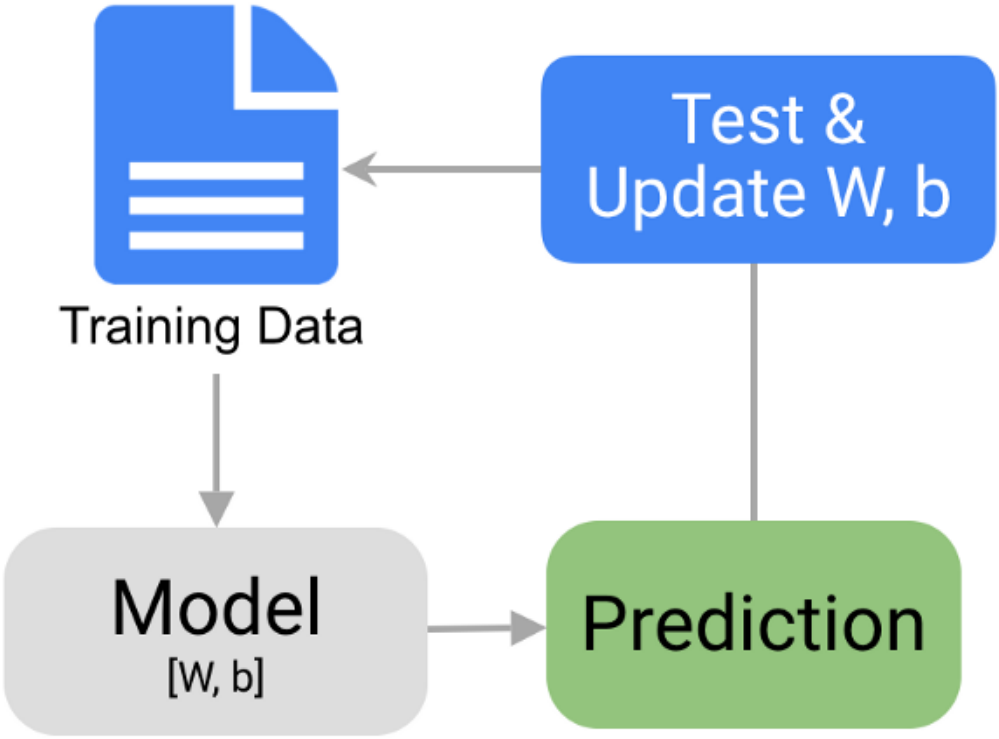


Figure: Machine Learning Modelling process

Once we are done with the data cleaning, we are going to split our data to train and test datasets. Training datasets are used to teach our machine learning models while test datasets are used for validation. It is used to check how our predictive variable performs.

In this phase we use various statistical techniques to identify patterns within the data.

4 steps in the Modelling Phase are:

a. First, we must select modelling techniques that we need to use for the prepared dataset.

b. Next, we must generate a test scenario to validate the quality and validity of the model.

c. Then, by using modelling tools we must prepare one or more models on the dataset.

d. Finally, these models need to be assessed by the project's stakeholders. That is to make sure that the models meet business initiatives.

The target variable for our task is " Credit Rating". Below is an example of how the variable is classified:

Credit Ratings*

		Moody's	S&P's	Fitch
Investment Grade	Strongest	Aaa	AAA	AAA
		Aa	AA	AA
		A	A	A
		Baa	BBB	BBB
Non Investment Grade		Ba	BB	BB
		B	B	B
		Caa	CCC	CCC
		Ca	CC	CC
		C	C	C
	Weakest	D	D	D

*These credit ratings are reflective of obligations with long-term maturities.

Figure: Credit Rating

For our modelling and overall analysis we are going to use the following packages such as sklearn, numpy, pandas etc.

By looking at our data and target variable we have decided to use the following algorithms:

a. Support vector machine
b. Random Forest
c. K neighbour classifier
d. Principal component analysis

SUPPORT VECTOR MACHINE

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. SVMs are more commonly used in classification problems and as such, this is what we will focus on in this post. (2)

SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes, as shown in the image below.

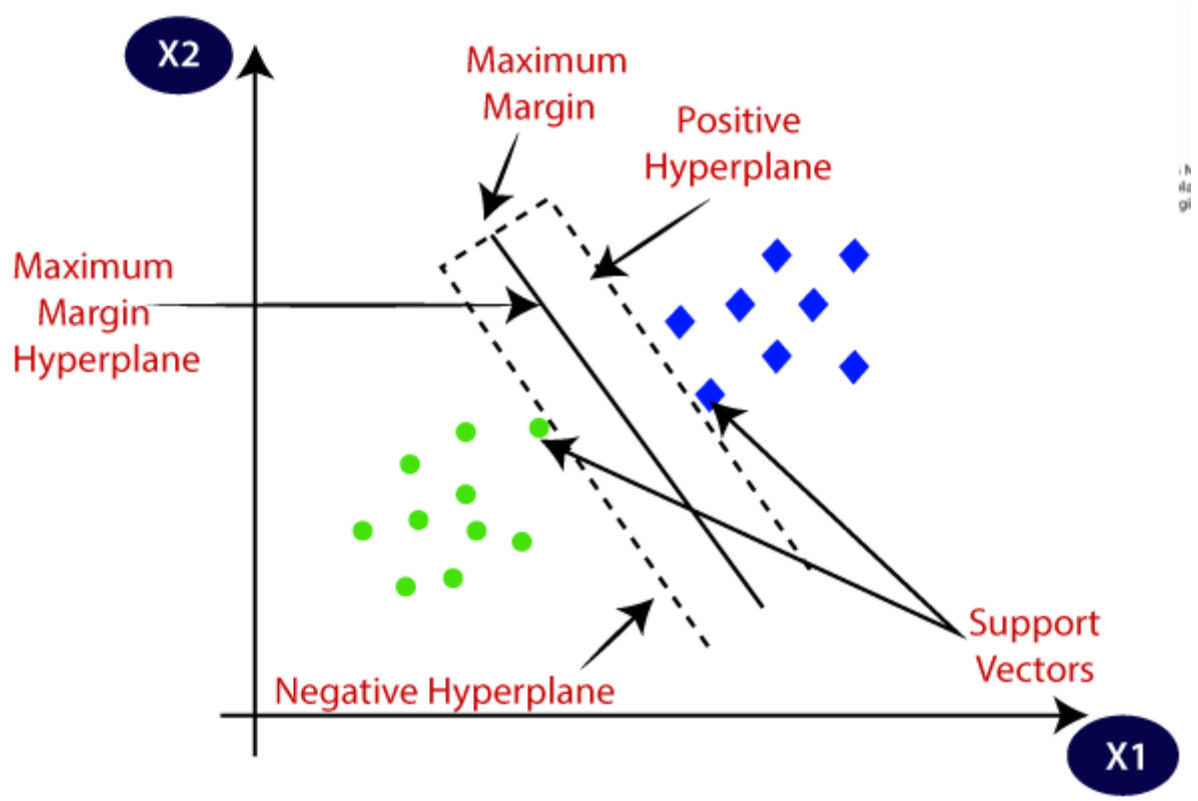


Figure: Support vector machine explained

Support vectors are the data points nearest to the hyperplane, the points of a data set that, if removed, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of a data set. (2)

The hyperplanes in the higher-dimensional space are defined as the set of points whose dot product with a vector in that space is constant, where such a set of vectors is an orthogonal (and thus minimal) set of vectors that defines a hyperplane.

RANDOM FOREST

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees. (3)

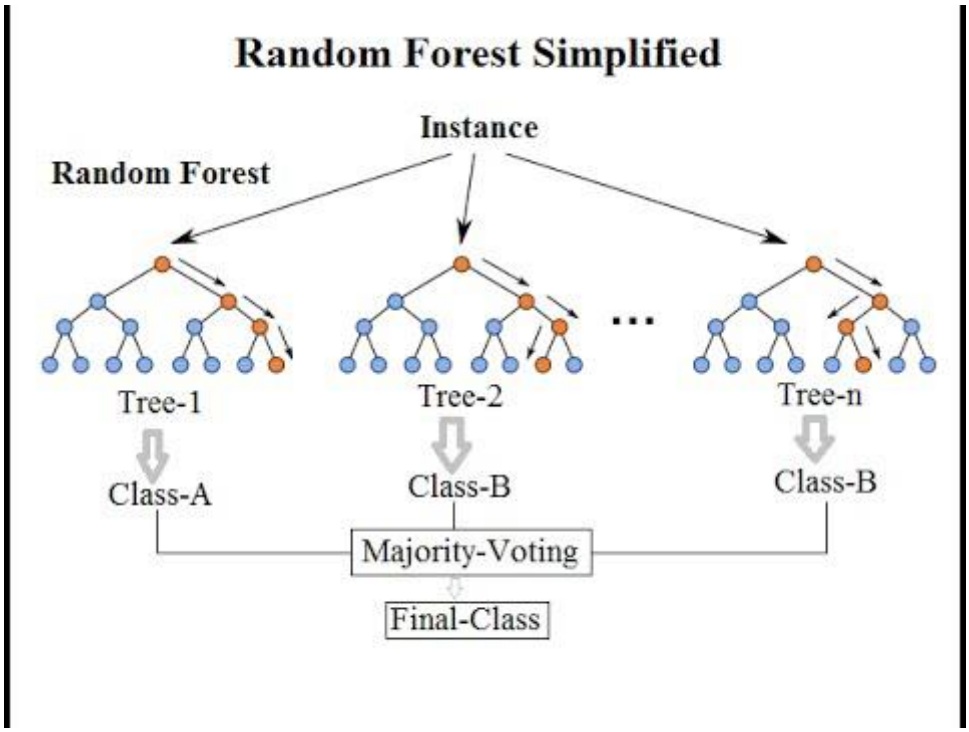


Figure: Random Forest Algorithm

Just to have a jest of decision trees, it split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. Each node in the decision tree works on a random subset of features to calculate the output. The random forest then combines the output of individual decision trees to generate the final output.

K Neighbours Classification

KNN algorithm is one of the simplest classification algorithm and it is one of the most used learning algorithms. KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. (4)

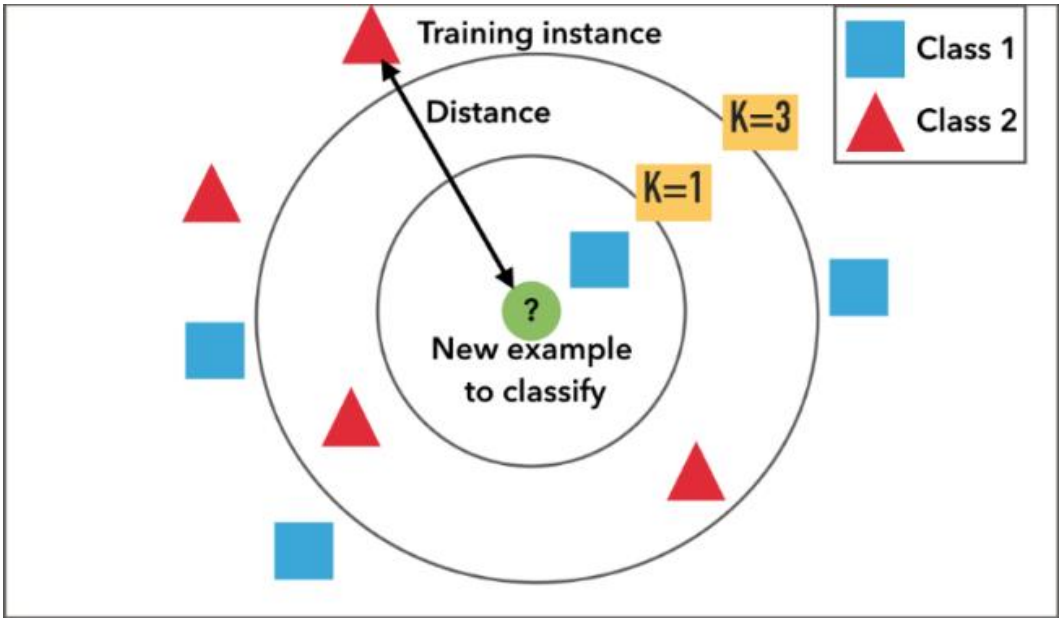


Figure: K neighbours' illustration

When we say a technique is non-parametric, it means that it does not make any assumptions on the underlying data distribution. In other words, the model structure is determined from the data. If you think about it, it's pretty useful, because in the "real world", most of the data does not obey the typical theoretical assumptions made (as in linear regression models, for example). Therefore, KNN could and probably should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution data. (4)

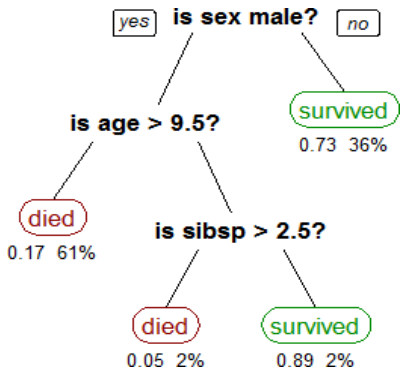


Figure: Decision Tree

PRINCIPAL COMPONENT ANALYSIS

Principle Component Analysis (PCA) is a measurable technique used to change over a set of somewhat associated factors to a lot of sprightly uncorrelated factors utilizing symmetrical changes. The factors yielded are in an arranged in such a way that the rest component has the most noteworthy fluctuation and pursued by different components in diminishing request. PCA is utilized for measurement decrease of the information which initially contained 98 highlights.

As we have seen in the data preparation, we have 98 features in our data set.

To consider all these features is a bit problematic so we shall try to reduce our data set using the principal component analysis.

The principle thought of principal component analysis (PCA) is to diminish the dimensionality of an informational collection comprising of numerous factors corresponded with one another, either vigorously or daintily, while holding the variety present in the dataset, up to the greatest degree.

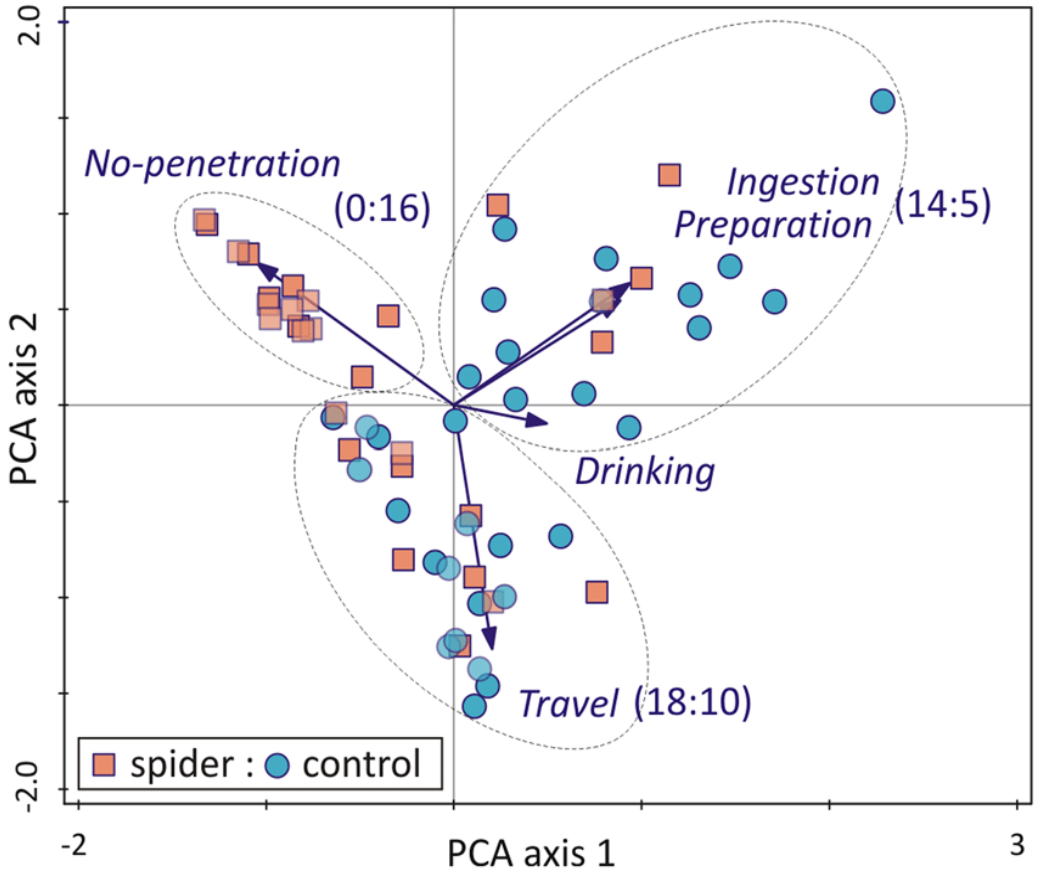


Figure: PCA illustration

Performance of the Models on the Dataset

Training Dataset:

The dataset consists of 1430 ratings of 400 companies considered over a period of 6 years from 2010 to 2016. A total of 98 features are used, where data from 49 financial quantities for the rating year and the year previous to it are used for each training example.

Cross Validation:

Cross Validation is used in each of the methods for optimizing the training of the model. These optimized models are constructed for SVM, K Neighbors Classifier, SVM and Random Forest which are then used for testing on the out-of sample date for accuracy. The technique of grid search is used for searching the best parameters for the models. Also, the generated results are reported after 7-fold cross validation for higher consistency.

Classification of Data:

The Credit ratings are categorized into 5 coarser credit categories (’AA’, ’A’, ’BBB’,’BB’, ’C’).

Rating	Number of Companies	%
AA	27	2.2
A	163	13.3
BBB	548	44.84
BB	246	20.13
C	238	19.47

Performance Measurement Metrics:

Different metrics are used in evaluating the performance of the models.

1. Precision Score

2. Recall Score

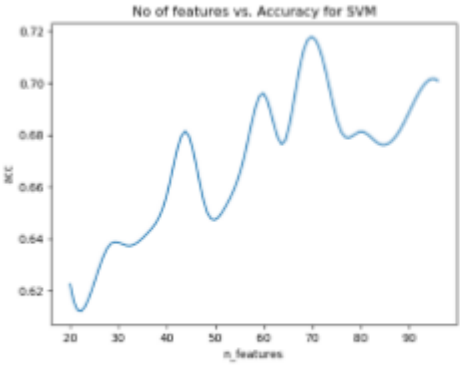
3. F1 Score

Feature Selection:

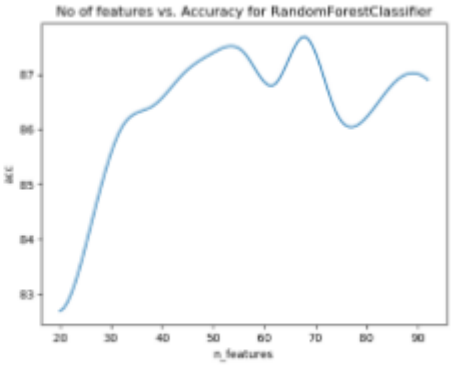
Univariate feature selection is performed based on univariate statistical tests to select the best features amongst all. Features are selected according to their percentile of the highest score.

Dimensional analysis:

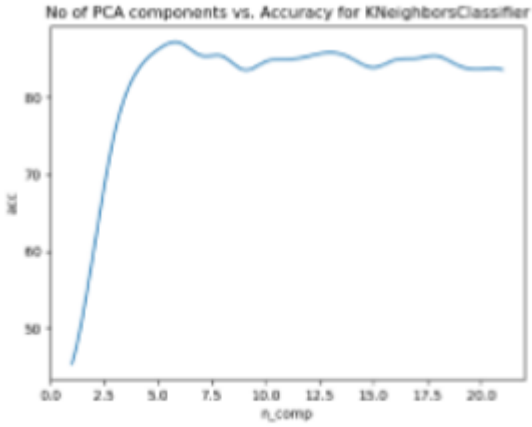
Model accuracy was checked for different dimensions of the training set for three different models. The output is shown below:



(a) SVM



(b) Random Forest Classifier



(c) KNeighbours Classifier

Principal Component Analysis:

1. Principal Component Analysis (PCA) is a statistical method used to convert a set of partially correlated variables to a set of linearly uncorrelated variables using orthogonal transformations.

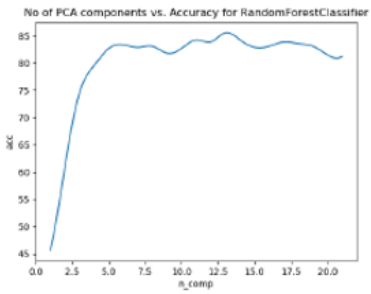
2. The variables outputted are in a sorted in such a way that the first component has the highest variance and followed by other components in decreasing order.

3. PCA is used for dimension reduction of the data which originally contained 98 features.

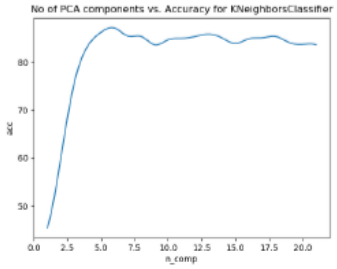
4. Model accuracies were checked for different number of PCA components for the four different learning algorithms.

Use of PCA for final results:

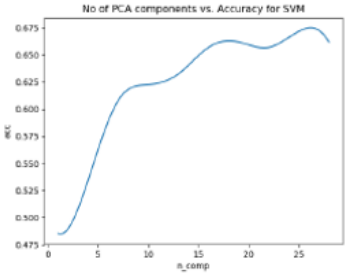
No significant improvement in accuracies was observed using PCA components. Hence, the original features were retained for all the subsequent learning algorithms.



(b) Random Forest Classifier



(a) KNeighbours Classifier



(a) SVM

Modelling Results:

We used different machine learning models to predict the credit ratings on the test

dataset. The following model accuracies were obtained:

As we can see maximum accuracy was achieved by using KNN, but RFC and SVC also show promising results.

	SVC	**RFC**	**KNC**
Accuracy	71.25	85.43	92.1

Conclusion and Future Scope

This project applies the techniques of SVM, Random Forest and KNeighbors Classifier to the problem of corporate credit rating prediction, and it is seen that these methods are indeed quite promising for application in this financial domain. This is one of the most crucial and time-consuming issues faced by the financial market.

The results obtained in this study clearly demonstrate high accuracy compared to the existing methods in the literature for the credit rating prediction problem. In particular, the results of K Neighbors Classifier is quite impressive and provoke further thought towards application.

The excellent performance of the instance-based learning model KNeighbors Classifier in this problem is an interesting topic to dive into, given the novelty of this approach.

In future, we can also use neural networks to enhance the performance of the prediction model and gain higher accuracies. Further study can be conducted to see how the algorithms scale with more data.

Bibliography

1. **corporatefinanceinstitute.** [Online] https://corporatefinanceinstitute.com/resources/knowledge/finance/credit-rating/#:~:text=Credit%20Score,rendered%20in%20the%20first%20place..

2. **KDnuggets.** support-vector-machines-simple-explanatio. *KDnuggets.com.* [Online] https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html.

3. **Wikipedia.** Random_forest. [Online] https://en.wikipedia.org/wiki/Random_forest.

4. **Bronshiein, Adi.** k-nearest neghibours alogorithm. [Online] 12 April 2017. https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7.