FINANCE AND RISK ANALYTICS

Table of Contents

Contents

Problem Statement

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. This project will cover the basic model building for predicting credit risk.

Data Description

The dataset consists of information from the financial statement of the companies for the previous year (2015). Also, information about the Networth of the company in the following year (2016) is provided which can be used to drive the labeled field.

Sample of the dataset:

	Co_Code	Co_Name	Networth Next Year	Equity Paid Up	Networth	Capital Employed	Total Debt	Gross Block	Net Working Capital	Current Assets	...	PBIDTM (%) [Latest]	PBITM (%) [Latest]	PBDTM (%) [Latest]	CPM (%) [Latest]	APATM (%) [Latest]	Debtors Velocity (Days)	Creditors Velocity (Days)
0	16974	Hind.Cables	-8021.60	419.36	-7027.48	-1007.24	5936.03	474.30	-1076.34	40.50	...	0.00	0.00	0.00	0.00	0.00	0	0
1	21214	Tata Tele. Mah.	-3986.19	1954.93	-2968.08	4458.20	7410.18	9070.86	-1098.88	486.86	...	-10.30	-39.74	-57.74	-57.74	-87.18	29	101
2	14852	ABG Shipyard	-3192.58	53.84	506.86	7714.68	6944.54	1281.54	4496.25	9097.64	...	-5279.14	-5516.98	-7780.25	-7723.67	-7961.51	97	558
3	2439	GTL	-3054.51	157.30	-623.49	2353.88	2326.05	1033.69	-2612.42	1034.12	...	-3.33	-7.21	-48.13	-47.70	-51.58	93	63
4	23505	Bharati Defence	-2967.36	50.30	-1070.83	4675.33	5740.90	1084.20	1836.23	4685.81	...	-295.55	-400.55	-845.88	379.79	274.79	3887	346

Q1.8 Build a Random Forest Model on Train Dataset. Also showcase your model building approach

Random Forest classifier from Sklearn is built on the dataset.

```
rfcl = RandomForestClassifier()

grid_search = GridSearchCV(estimator = rfcl, param_grid = param_grid)
```

Random Forest Classifier is built using hyper parameter tuning. Hyper parameters tuning for this model are done using GRidSearchCV method.

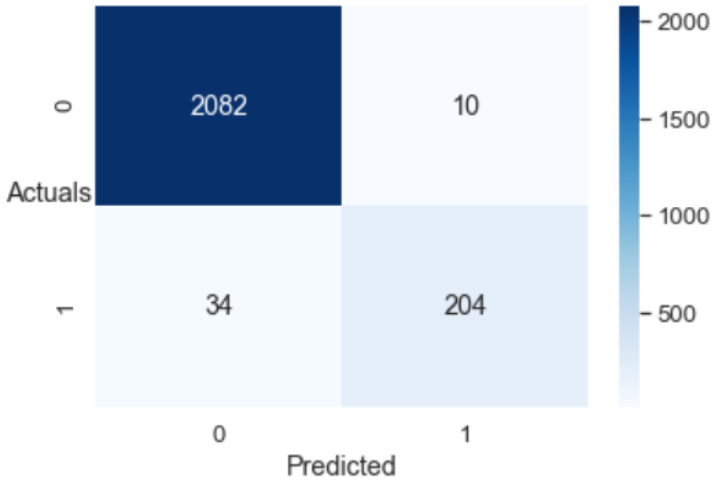
```
GridSearchCV(estimator=RandomForestClassifier(),
             param_grid={'max_depth': [3, 5, 7],
                         'min_samples_leaf': [5, 10, 15],
                         'min_samples_split': [15, 30, 45],
                         'n_estimators': [25, 50]})
```

Performance Metrics of Random Forest Model on Training Set

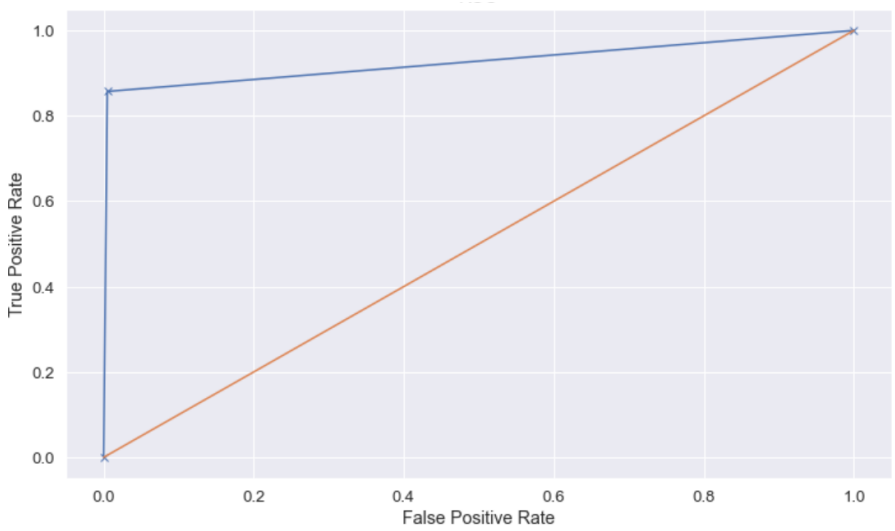
Classification Report

	precision	recall	f1-score	support
0	0.98	1.00	0.99	2092
1	0.95	0.86	0.90	238
accuracy			0.98	2330
macro avg	0.97	0.93	0.95	2330
weighted avg	0.98	0.98	0.98	2330

Confusion Matrix



AUC/ROC Plot




```
Area under Curve is 0.926181371210052
```

Q1.9 Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model

Now Random Forest Model is validated on the dataset.

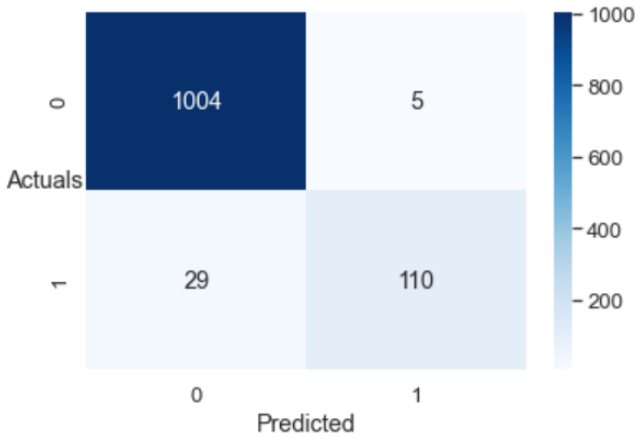
Performance Metrics on Testing Dataset

Classification Report

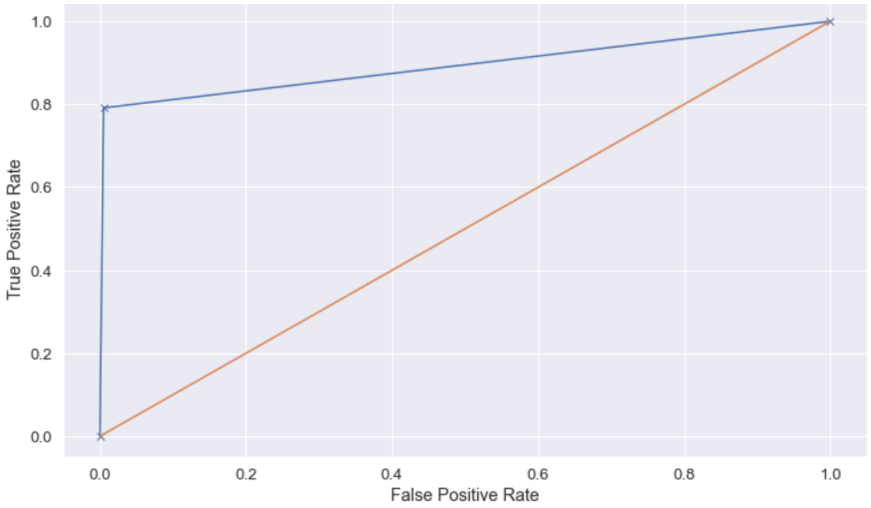
	precision	recall	f1-score	support
0	0.97	1.00	0.98	1009
1	0.96	0.79	0.87	139
accuracy			0.97	1148
macro avg	0.96	0.89	0.92	1148
weighted avg	0.97	0.97	0.97	1148

The accuracy score of Random forest Model on Test Dataset is 97%.

Confusion Matrix



ROC Plot and AUC score



Area under Curve is 0.8932057525436539

AUC Score is 0.89 for this model. Higher the AUC score, better the model. Random Forest model's accuracy score is higher when compared to Logistic Regression's accuracy score. When compared to Logostoc Regression, random forest is a better model.

Q1.10 Build a LDA Model on Train Dataset. Also showcase your model building approach

LDA is built using sklearn.

Note that scaling does not provide any significant difference to results, therefore it is not done in this case.

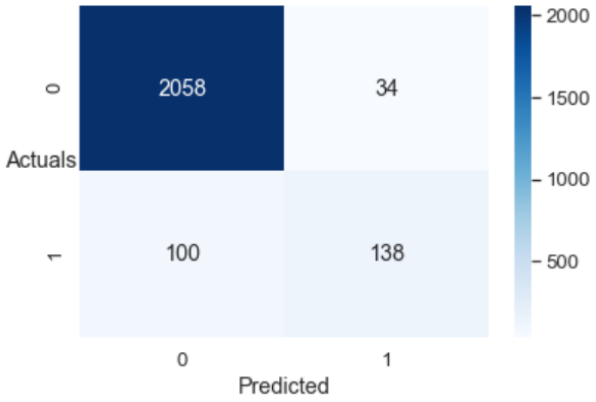
```
lda_model = LDA.fit(X_train, y_train)
```

Performance metrics on Training Dataset

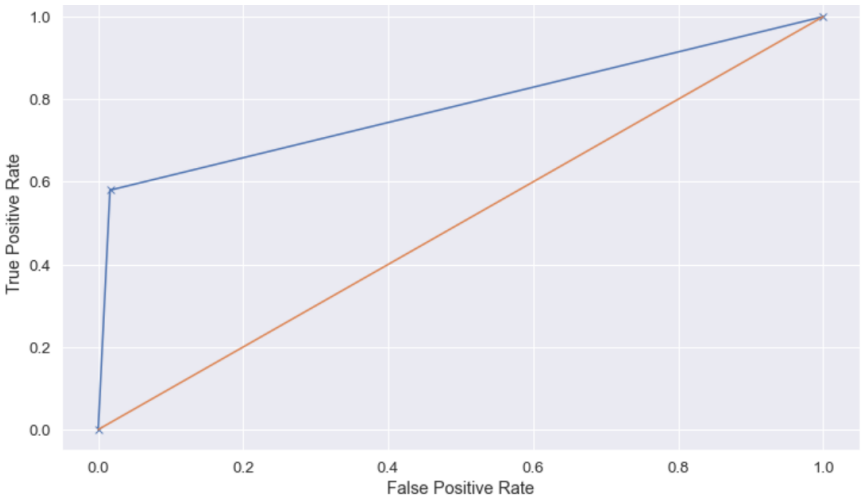
Classification Report

	precision	recall	f1-score	support
0	0.95	0.98	0.97	2092
1	0.80	0.58	0.67	238
accuracy			0.94	2330
macro avg	0.88	0.78	0.82	2330
weighted avg	0.94	0.94	0.94	2330

Confusion Matrix



AUC Score/ROC plot




```
Area under Curve is 0.7817897713578741
```

Q1.11 Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model

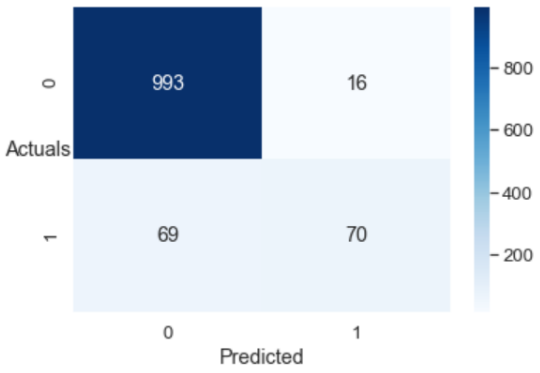
Linear Discriminant Analysis is validated on the Test dataset.

Performance Metrics on Test Dataset

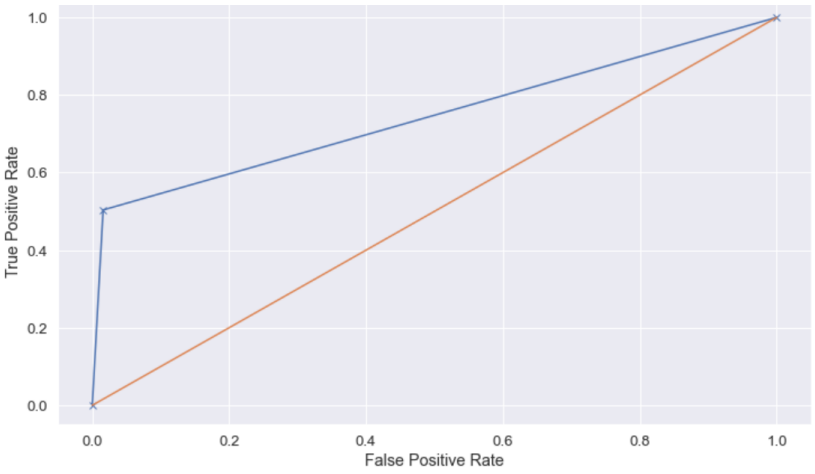
Classification Report

	precision	recall	f1-score	support
0	0.94	0.98	0.96	1009
1	0.81	0.50	0.62	139
accuracy			0.93	1148
macro avg	0.87	0.74	0.79	1148
weighted avg	0.92	0.93	0.92	1148

Confusion Matrix



AUC Score/ROC plot



Area under Curve is 0.7438699189310594

The accuracy score for this model on testing set is 93%. Although, AUC Score is low compared to the other two models (Logistic Regression and Random Forest Model).

Q 1.12 Compare the performances of Logistics, Radom Forest and LDA models (include ROC Curve)

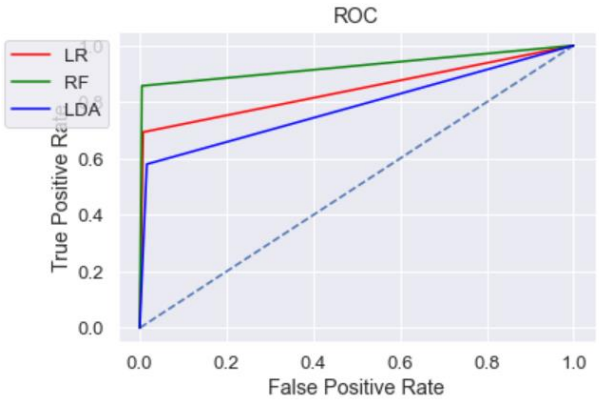
Performance Metrics of Models on both Training and Testing Datasets.

	LR Train	LR Test	RF Train	RF Test	LDA Train	LDA Test
Accuracy	0.96	0.89	0.98	0.97	0.94	0.93
AUC	0.84	0.92	0.93	0.89	0.78	0.74
Recall	0.69	0.97	0.86	0.79	0.58	0.50
Precision	0.91	0.52	0.95	0.96	0.80	0.81
F1 Score	0.79	0.68	0.90	0.87	0.67	0.62

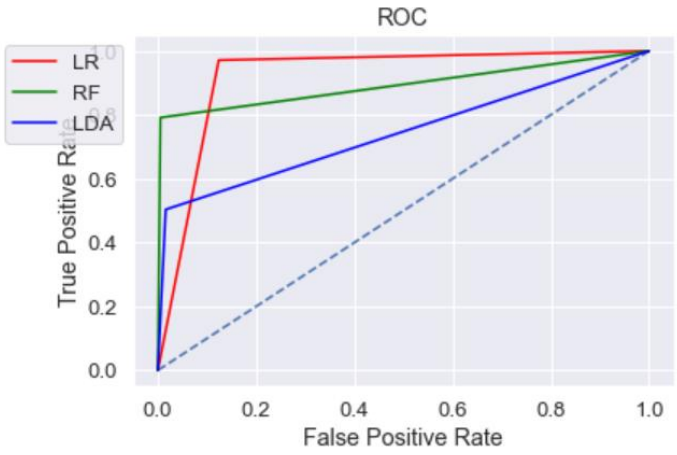
As observed, Accuracy score of Random Forest Model is highest with 97% followed by LDA and Logistic Regression.

Random Forest is the best performing model.

ROC Plots of Models on Training Dataset



ROC Plots of Models on Testing Dataset



Logistic Regression's line is slightly higher than the Random forest. Although, Precision and F1 scores of the logistic regression is lower when compared to Random Forest model.

Therefore, Random Forest model is the best model amongst all the models with 97.01% accuracy, 96% precision, F1-score with 0.87 and recall score of 0.79.

Q **1.13 State Recommendations from the above models.**

Reccomendations are given looking at the coefficient values of the variables.

- Book_Value_Adj_Unit_Curr has p-value of 0.000 which suggests that there is a high chance of default if this variable is decreased.

- Current_Ratio_Latest is inversely proportional to the customer defaulting. If the companies' ability to pay in short term dues are higher, then lower is the chance of default and having a negative networth for the company.

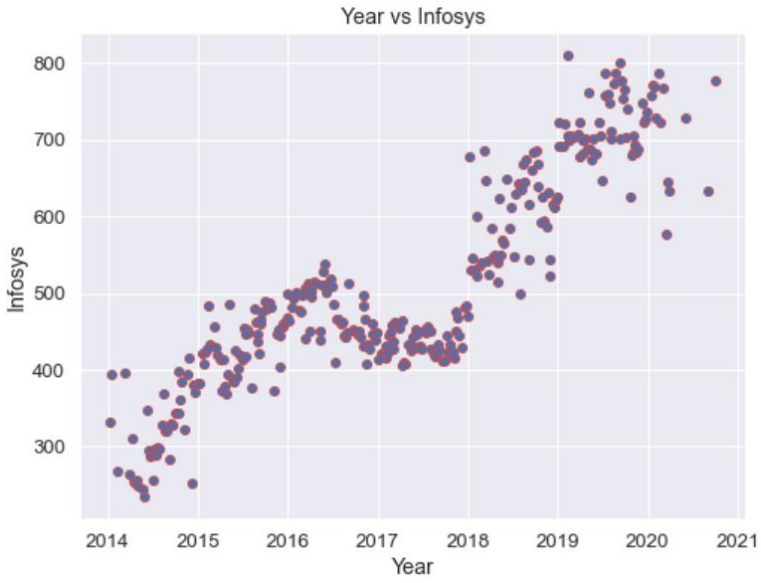
- PBDTM_perc_Latest and Total Debt are other two important variables that need to be considered while building a model.

- Random Forest and Logistic regression models can be built using SMOTE for better results.

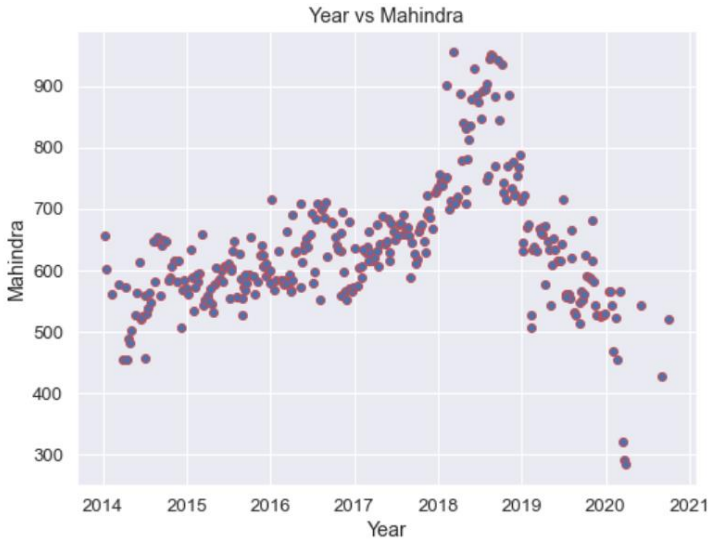
Q 2.1 Draw Stock Price Graph(Stock Price vs Time) for any 2 given stocks with inference

The dataset contains 6 years of information (weekly stock information) on the stock prices of 10 different Indian Stocks.

Infosys Stock Price vs Time



Mahindra Stock Price vs Time



Inference: Stock Price for Infosys is consistently increasing over time. But Stock price for Mahindra dropped significantly from 2018.

Q 2.2 Calculate Returns for all stocks with inference

Returns for all the stocks i.e. difference of log of price at t and the log of price at t-1 are shown below.

```
stock_returns = np.log(stock_prices.drop(['Date','dates'],axis=1)).diff(axis = 0, periods = 1)
```

	Infosys	Indian_Hotel	Mahindra_&_Mahindra	Axis_Bank	SAIL	Shree_Cement	Sun_Pharma	Jindal_Steel	Idea_Vodaf
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
1	-0.026873	-0.014599	0.006572	0.048247	0.028988	0.032831	0.094491	-0.065882	0.011
2	-0.011742	0.000000	-0.008772	-0.021979	-0.028988	-0.013888	-0.004930	0.000000	-0.011
3	-0.003945	0.000000	0.072218	0.047025	0.000000	0.007583	-0.004955	-0.018084	0.000
4	0.011788	-0.045120	-0.012371	-0.003540	-0.076373	-0.019515	0.011523	-0.140857	-0.049

Q 2.3 Calculate Stock Means and Standard Deviation for all stocks with inference

Stock Means:

Average returns that the stock is making on a week to week basis

```
Shree_Cement           0.003681
Infosys                0.002794
Axis_Bank              0.001167
Indian_Hotel           0.000266
Sun_Pharma            -0.001455
Mahindra_&_Mahindra   -0.001506
SAIL                  -0.003463
Jindal_Steel          -0.004123
Jet_Airways           -0.009548
Idea_Vodafone         -0.010608
dtype: float64
```

Inference: Shree company has the highest returns and Idea_Vodafone has the lowest returns.

Stock Standard Deviation :

It is a measure of volatility meaning the more a stock's returns vary from the stock's average return, the more volatile the stock

```
Idea_Vodafone          0.104315
Jet_Airways            0.097972
Jindal_Steel           0.075108
SAIL                   0.062188
Indian_Hotel           0.047131
Axis_Bank              0.045828
Sun_Pharma             0.045033
Mahindra_&_Mahindra    0.040169
Shree_Cement           0.039917
Infosys                0.035070
dtype: float64
```

Inference: Idea_Vodafone has the highest risk factor while Infosys is the least risky investment option

Q 2.4 **Draw a plot of Stock Means vs Standard Deviation and state your inference**

Inference: Stocks that are on upper left side have high volatility and low returns, while the stocks on the bottom right indicate low volatility and high returns.

Q 2.5 **Conclusion and Recommendations**

- Stock with a lower mean & higher standard deviation do not play a role in a portfolio that has competing stock with more returns & less risk. Thus we can keep the following stocks out of our portfolio which have negetive returns with high volatality.

    a) Idea_Vodafone
    b) Jet_Airways
    c) Jindal_Steel
    d) SAIL

- Shree Cements seems like a good option for high returns and low risk.
- Medium Returns with low risks, Infosys can be considered.
- Axis bank and Indian Hotels are options for low returns and low risk.
- Stock means vs standard deviation plot is used to recommend. More volatile stock might give short term gains but might not be a good investment in long term.