PROJECT REPORT

DSBA

By
Siddhanth Gupta

greatlearning
Learning for Life

ROC Curve

The **ROC curve** for the various models shows that the **training and test set results are similar**, with the model exhibiting **high overall measures**. This indicates that the model is performing well in predicting the default status.

Net worth is identified as the most important variable for predicting the likelihood of default, which further validates its significance in the analysis.

Building an LDA Model on Train Dataset

Linear Discriminant Analysis (LDA) and Logistic Regression are widely used for analyzing categorical outcome variables. Both techniques are effective for developing linear classification models, but LDA is preferred when the outcome variable has more than two groups. LDA also assumes a normal distribution of the predictors, making it useful for data with normally distributed features.

	Date	Infosys	Indian Hotel	Mahindra & Mahindra	Axis Bank	SAIL	Shree Cement	Sun Pharma	Jindal Steel	Idea Vodafone	Jet Airways
0	31-03-2014	264	69	455	263	68	5543	555	298	83	278
1	07-04-2014	257	68	458	276	70	5728	610	279	84	303
2	14-04-2014	254	68	454	270	68	5649	607	279	83	280
3	21-04-2014	253	68	488	283	68	5692	604	274	83	282
4	28-04-2014	256	65	482	282	63	5582	611	238	79	243

The **coefficients from the LDA model** indicate the importance of different variables in predicting default. The highest importance feature identified by LDA is **BookValueAdjUnitCurr** with a coefficient of **-2.0044213**.

Validating the LDA Model on the Test Dataset

The model was validated using the **test dataset** with the following performance matrices:

- **Confusion Matrix**
- **Classification Report**
- **ROC Curve**

Training and test set results were similar, and the overall measures were high, reinforcing that the LDA model is a good model for predicting defaults. As with other models, **BookValueAdjUnitCurr** emerged as the most important variable for predicting default status.

Comparing the Performances of Logistic Regression, Random Forest, and LDA Models

Comparison of the performance metrics from the three models:

Metric	Logistic Train	Logistic Test	LDA Train	LDA Test	Random Forest Train	Random Forest Test
Accuracy	0.94	0.93	0.92	0.90	0.97	0.96
AUC	0.934	0.946	0.932	0.941	0.87	0.87
Recall	0.49	0.58	0.36	0.31	0.76	0.76
Precision	0.83	0.83	0.78	0.75	0.93	0.93
F1 Score	0.62	0.68	0.49	0.44	0.83	0.84

• **Accuracy**: **Random Forest** has the highest accuracy (**0.96**) on the test set.

• **AUC**: **Logistic Regression** has the highest AUC value (**0.946**), while **Random Forest** has the lowest AUC (**0.87**).

• **Recall**: **Random Forest** achieves the highest recall (**0.76**), and **LDA** has the lowest (**0.31**).

• **Precision**: **Random Forest** has the highest precision (**0.93**), while **LDA** has the lowest (**0.75**).

• **F1 Score**: **Random Forest** has the highest F1 score (**0.84**), while **LDA** has the lowest (**0.44**).

In conclusion, **Random Forest** demonstrates slightly better performance than **Logistic Regression** and **LDA** in most metrics. All three models are stable enough for future predictions.

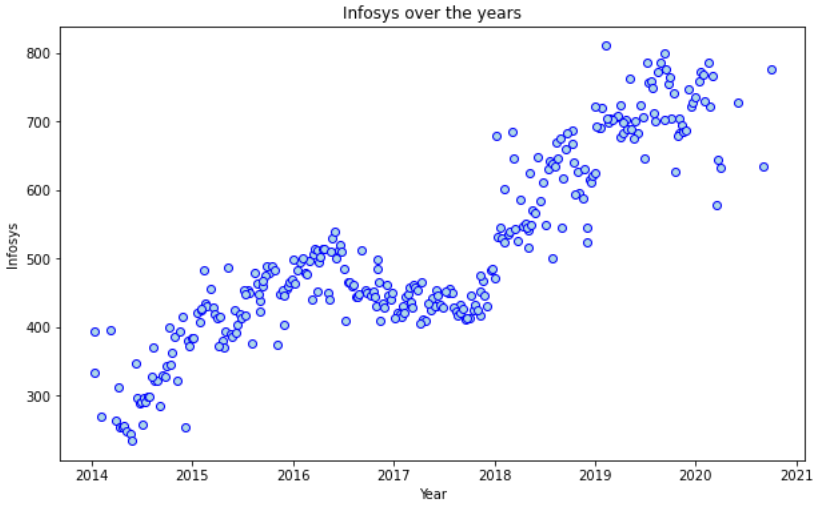
The model shows that Random Forest is an effective algorithm for predicting company defaults due to its performance and lower preprocessing requirements. Hyperparameter tuning for Random Forest is also easier compared to other models. This makes it the preferred choice for predicting the likelihood of default.

The **BookValueAdjUnitCurr** variable is consistently identified as the most crucial feature for predicting default. Therefore, it should be closely monitored for better prediction accuracy. The model can help businesses and investors assess the likelihood of default by analyzing financial variables in the dataset.

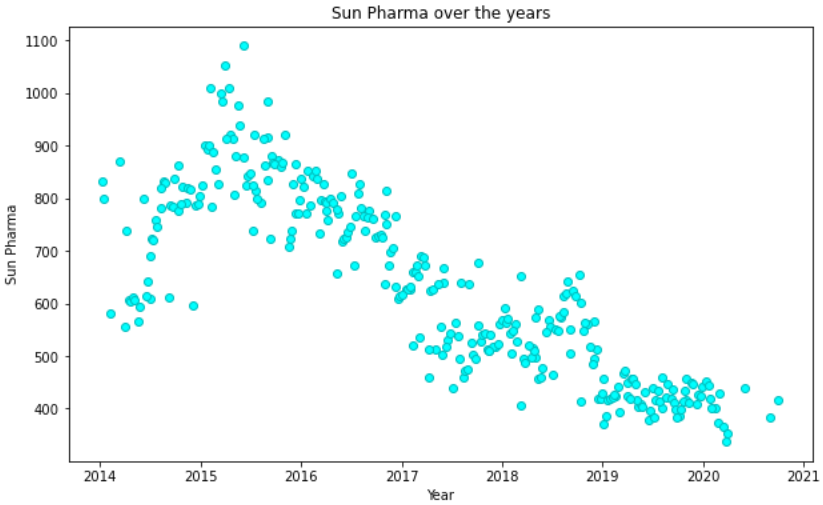
2. Stock Price Analysis

Stock Price Graph (Stock Price vs Time)

• **Infosys**: The stock price shows an increasing trend, with a significant rise in price over the period analyzed.



• **Sun Pharma**: The stock price demonstrates a decreasing trend, with a notable drop in price during the same period.



Calculating Returns for All Stocks

Returns represent the difference in stock price between two consecutive weeks.

• A **negative return** indicates a **decrease** in stock price compared to the previous week, while a **positive return** indicates an **increase**.

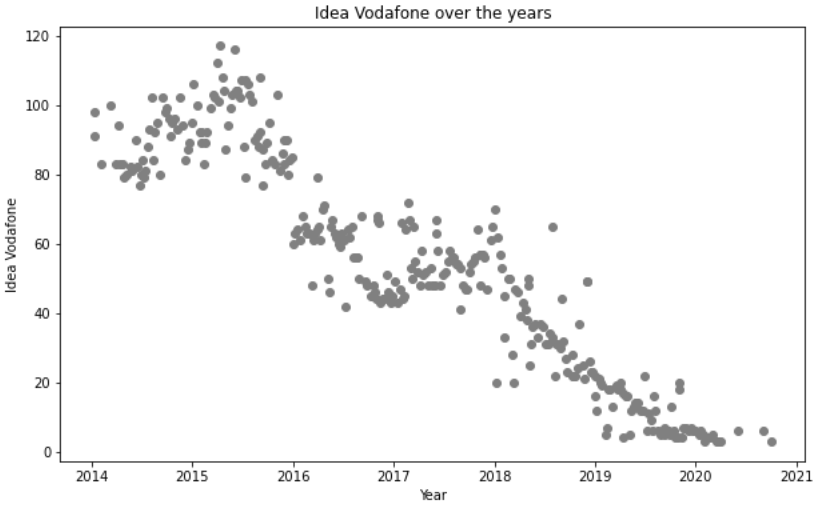
Calculating Stock Means and Standard Deviation

• **Stock Means**: The average returns of the stock on a week-to-week basis.

• **Shree Cement** has the highest mean return, and **Jet Airways** has the lowest.

• **Stock Standard Deviation**: A measure of volatility, indicating how much the stock's returns vary from its average return.

• **Idea Vodafone** has the highest volatility, and **Infosys** has the lowest.



2.4) Stock Means vs Standard Deviation Plot

The plot reveals that stocks with **higher average returns** tend to have **lower volatility**. In contrast, stocks with **lower returns** often exhibit **higher volatility**.

Conclusion and Recommendations

The following stocks show positive average returns:

• **Infosys**: 0.002794

• **Indian Hotel**: 0.000266

• **Axis Bank**: 0.001167

• **Shree Cement**: 0.003681

	Average	Volatility
Infosys	0.002794	0.035070
Shree Cement	0.003681	0.039917
Mahindra & Mahindra	-0.001506	0.040169
Sun Pharma	-0.001455	0.045033
Axis Bank	0.001167	0.045828
Indian Hotel	0.000266	0.047131
SAIL	-0.003463	0.062188
Jindal Steel	-0.004123	0.075108
Jet Airways	-0.009548	0.097972
Idea Vodafone	-0.010608	0.104315

Among these, **Infosys** and **Shree Cement** offer **higher returns with lower volatility**, making them the most attractive options for investment. Stocks with **lower returns and higher volatility** should be avoided as they do not contribute positively to a diversified portfolio. Therefore, **Infosys** and **Shree Cement** are recommended for stable investment with favorable risk-return profiles.