Project Members:

- Siddhant Angne
- Aniket Patil
- Atharva Teli

Project Name: Company Bankruptcy Prediction

Data Source: https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction

CSV Files used:

- data.csv

Number of Rows: 6819

Number of Columns: 96

Introduction:

The Company Bankruptcy Prediction problem is a regression problem, and we are using Decision Forest Regression Algorithm, Boosted Decision Tree Regression Algorithm and Neutral Network Regression Algorithm to predict what will cause bankruptcy to the company. Based on the data in columns of the dataset we can identify what causes the company to go bankrupt.

Member Name	Algorithm
Siddhant Angne	Boosted Decision Tree Regression Algorithm
Aniket Patil	Decision Forest Regression Algorithm
Atharva Teli	Neutral Network Regression Algorithm

Characteristics of Dataset:

data.csv

Columns	Description	Data Types
Bankrupt?	Bankrupt?: Class label	Integer
ROA(C) before interest and depreciation before interest	ROA(C) before interest and depreciation before interest: Return On Total Assets(C)	Decimal
ROA(A) before interest and % after tax	ROA(A) before interest and % after tax: Return On Total Assets(A)	Decimal
ROA(B) before interest and depreciation after tax	ROA(B) before interest and depreciation after tax: Return On Total Assets(B)	Decimal
Operating Gross Margin	Operating Gross Margin: Gross Profit/Net Sales	Decimal
Realized Sales Gross Margin	Realized Sales Gross Margin: Realized Gross Profit/Net Sales	Decimal
Operating Profit Rate	Operating Profit Rate: Operating Income/Net Sales	Decimal
Pre-tax net Interest Rate	Pre-tax net Interest Rate: Pre-Tax Income/Net Sales	Decimal
After-tax net Interest Rate	After-tax net Interest Rate: Net Income/Net Sales	Decimal

Non-industry income and expenditure/revenue	Non-industry income and expenditure/revenue: Net Non-operating Income Ratio	Decimal
Continuous interest rate (after tax)	Continuous interest rate (after tax): Net Income-Exclude Disposal Gain or Loss/Net Sales	Decimal
Operating Expense Rate	Operating Expense Rate: Operating Expenses/Net Sales	Decimal
Research and development expense rate	Research and development expense rate: (Research and Development Expenses)/Net Sales	Whole Number
Cash flow rate	Cash flow rate: Cash Flow from Operating/Current Liabilities	Decimal
Interest-bearing debt interest rate	Interest-bearing debt interest rate: Interest-bearing Debt/Equity	Decimal
Tax rate (A)	Tax rate (A): Effective Tax Rate	Decimal
Net Value Per Share (B)	Net Value Per Share (B): Book Value Per Share(B)	Decimal
Net Value Per Share (A)	Net Value Per Share (A): Book Value Per Share(A)	Decimal
Net Value Per Share (C)	Net Value Per Share (C): Book Value Per Share(C)	Decimal
Persistent EPS in the Last Four Seasons	Persistent EPS in the Last Four Seasons: EPS-Net Income	Decimal
Cash Flow Per Share	Cash Flow Per Share	Decimal
Revenue Per Share (Yuan Â¥)	Revenue Per Share (Yuan ¥): Sales Per Share	Decimal
Operating Profit Per Share (Yuan Â¥)	Operating Profit Per Share (Yuan ¥): Operating Income Per Share	Decimal
Per Share Net profit before tax (Yuan Â¥)	Per Share Net profit before tax (Yuan ¥): Pretax Income Per Share	Decimal
Realized Sales Gross Profit Growth Rate	Realized Sales Gross Profit Growth Rate	Decimal
Operating Profit Growth Rate	Operating Profit Growth Rate: Operating Income Growth	Decimal
After-tax Net Profit Growth Rate	After-tax Net Profit Growth Rate: Net Income Growth	Decimal
Regular Net Profit Growth Rate	Regular Net Profit Growth Rate: Continuing Operating Income after Tax Growth	Decimal
Continuous Net Profit Growth Rate	Continuous Net Profit Growth Rate: Net Income-Excluding Disposal Gain or Loss Growth	Decimal
Total Asset Growth Rate	Total Asset Growth Rate: Total Asset Growth	Whole Number
Net Value Growth Rate	Net Value Growth Rate: Total Equity Growth	Decimal
Total Asset Return Growth Rate Ratio	Total Asset Return Growth Rate Ratio: Return on Total Asset Growth	Decimal
Cash Reinvestment %	Cash Reinvestment %: Cash Reinvestment Ratio	Decimal
Current Ratio	Current Ratio	Decimal
Quick Ratio	Quick Ratio: Acid Test	Decimal

Interest Expense Ratio	Interest Expense Ratio: Interest Expenses/Total Revenue	Decimal
Total debt/Total net worth	Total debt/Total net worth: Total Liability/Equity Ratio	Decimal
Debt ratio %	Debt ratio %: Liability/Total Assets	Decimal
Net worth/Assets	Net worth/Assets: Equity/Total Assets	Decimal
Long-term fund suitability ratio (A)	Long-term fund suitability ratio (A): (Long-term Liability+Equity)/Fixed Assets	Decimal
Borrowing dependency	Borrowing dependency: Cost of Interest-bearing Debt	Decimal
Contingent liabilities/Net worth	Contingent liabilities/Net worth: Contingent Liability/Equity	Decimal
Operating profit/Paid-in capital	Operating profit/Paid-in capital: Operating Income/Capital	Decimal
Net profit before tax/Paid-in capital	Net profit before tax/Paid-in capital: Pretax Income/Capital	Decimal
Inventory and accounts receivable/Net value	Inventory and accounts receivable/Net value: (Inventory+Accounts Receivables)/Equity	Decimal
Total Asset Turnover	Total Asset Turnover	Decimal
Accounts Receivable Turnover	Accounts Receivable Turnover	Decimal
Average Collection Days	Average Collection Days: Days Receivable Outstanding	Decimal
Inventory Turnover Rate (times)	Inventory Turnover Rate (times)	Decimal
Fixed Assets Turnover Frequency	Fixed Assets Turnover Frequency	Decimal
Net Worth Turnover Rate (times)	Net Worth Turnover Rate (times): Equity Turnover	Decimal
Revenue per person	Revenue per person: Sales Per Employee	Decimal
Operating profit per person	Operating profit per person: Operation Income Per Employee	Decimal
Allocation rate per person	Allocation rate per person: Fixed Assets Per Employee	Decimal
Working Capital to Total Assets	Working Capital to Total Assets	Decimal
Quick Assets/Total Assets	Quick Assets/Total Assets	Decimal
Current Assets/Total Assets	Current Assets/Total Assets	Decimal
Cash/Total Assets	Cash/Total Assets	Decimal
Quick Assets/Current Liability	Quick Assets/Current Liability	Decimal
Cash/Current Liability	Cash/Current Liability	Decimal
Current Liability to Assets	Current Liability to Assets	Decimal
Operating Funds to Liability	Operating Funds to Liability	Decimal
Inventory/Working Capital	Inventory/Working Capital	Decimal
Inventory/Current Liability	Inventory/Current Liability	Decimal

Current Liabilities/Liability	Current Liabilities/Liability	Decimal
Working Capital/Equity	Working Capital/Equity	Decimal
Current Liabilities/Equity	Current Liabilities/Equity	Decimal
Long-term Liability to Current Assets	Long-term Liability to Current Assets	Decimal
Retained Earnings to Total Assets	Retained Earnings to Total Assets	Decimal
Total income/Total expense	Total income/Total expense	Decimal
Total expense/Assets	Total expense/Assets	Decimal
Current Asset Turnover Rate	Current Asset Turnover Rate: Current Assets to Sales	Decimal
Quick Asset Turnover Rate	Quick Asset Turnover Rate: Quick Assets to Sales	Decimal
Working capitcal Turnover Rate	Working capitcal Turnover Rate: Working Capital to Sales	Decimal
Cash Turnover Rate	Cash Turnover Rate: Cash to Sales	Decimal
Cash Flow to Sales	Cash Flow to Sales	Decimal
Fixed Assets to Assets	Fixed Assets to Assets	Decimal
Current Liability to Liability	Current Liability to Liability	Decimal
Current Liability to Equity	Current Liability to Equity	Decimal
Equity to Long-term Liability	Equity to Long-term Liability	Decimal
Cash Flow to Total Assets	Cash Flow to Total Assets	Decimal
Cash Flow to Liability	Cash Flow to Liability	Decimal
CFO to Assets	CFO to Assets	Decimal
Cash Flow to Equity	Cash Flow to Equity	Decimal
Current Liability to Current Assets	Current Liability to Current Assets	Decimal
Liability-Assets Flag	Liability-Assets Flag: 1 if Total Liability exceeds Total Assets, 0 otherwise	Whole Number
Net Income to Total Assets	Net Income to Total Assets	Decimal
Total assets to GNP price	Total assets to GNP price	Decimal
No-credit Interval	No-credit Interval	Decimal
Gross Profit to Sales	Gross Profit to Sales	Decimal
Net Income to Stockholder's Equity	Net Income to Stockholder's Equity	Decimal
Liability to Equity	Liability to Equity	Decimal
Degree of Financial Leverage (DFL)	Degree of Financial Leverage (DFL)	Decimal
Interest Coverage Ratio (Interest expense to EBIT)	Interest Coverage Ratio (Interest expense to EBIT)	Decimal
Net Income Flag	Net Income Flag: 1 if Net Income is Negative for the last two years, 0 otherwise	Whole Number
Equity to Liability	Equity to Liability	Decimal

Descriptive statistics of the dataset:

The respective data is included in summarize.csv for this content.

Train a machine learning model for each problem and report the process in the deliverable:

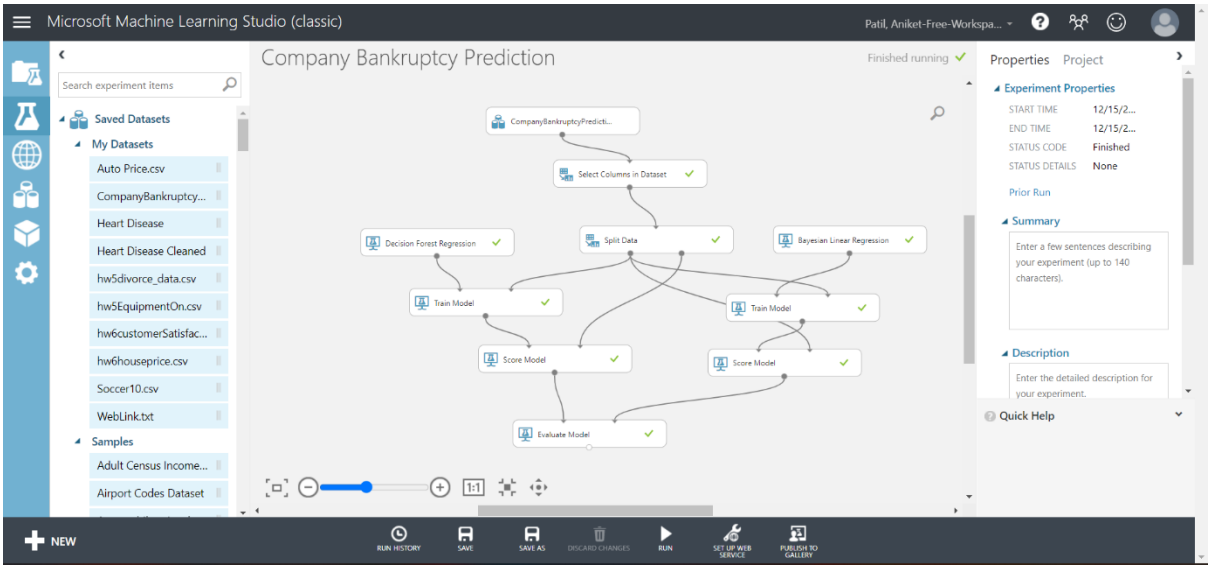
Model 1:

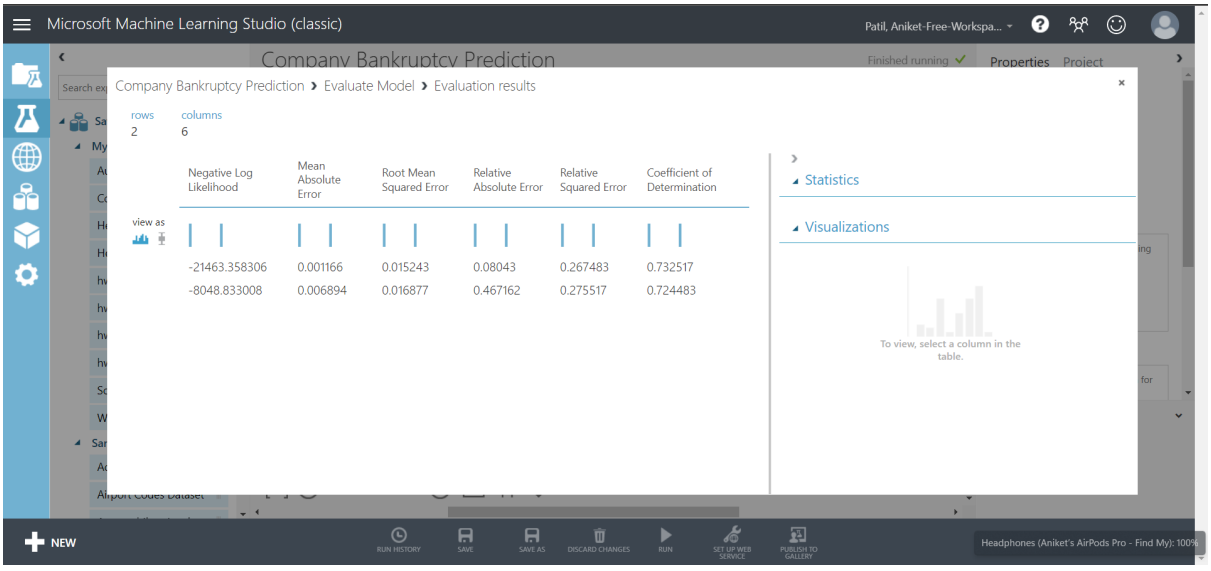
Member Name: Aniket Patil

Algorithm Chosen for training the model: Decision Forest Regression and Bayesian Linear Regression

Pre-processing: Excluded Liability-Assets Flag, Net Income Flag

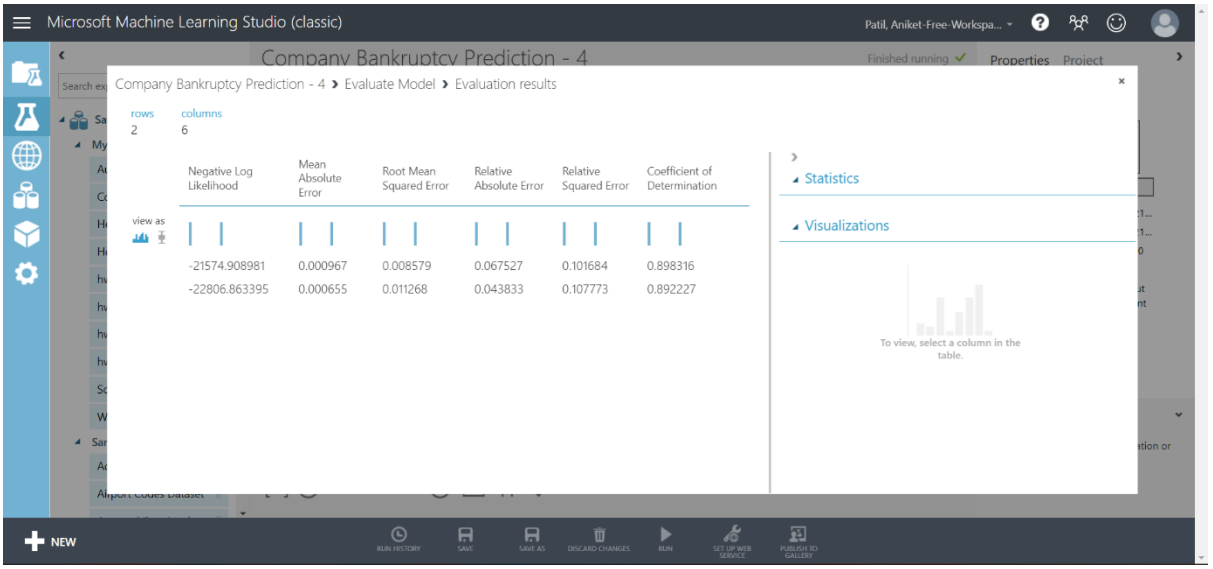
Best Algorithm: Decision Forest Algorithm





Performance of Decision Forest Algorithm is 73% which is greater than the performance of Bayesian Linear Regression Algorithm which is 72%, hence Decision Forest Algorithm is chosen to perform the Company Bankruptcy Prediction.

For the chosen algorithm, evaluate the degree of underfitting/overfitting problem by comparing the performance of the model for the scored training dataset with the performance of the model using the scored testing dataset. In the report, present model performance metrics and discuss underfitting/overfitting.



Since there is very negligible deviation between training and testing dataset it is qualifies for best fit dataset.

After applying the Tune Model Hyperparameters the performance went from best fit to under fitting and parameter range used as follows

Number of decision trees: 8, 16, 128

Maximum depth of the decision trees: 32, 45, 65

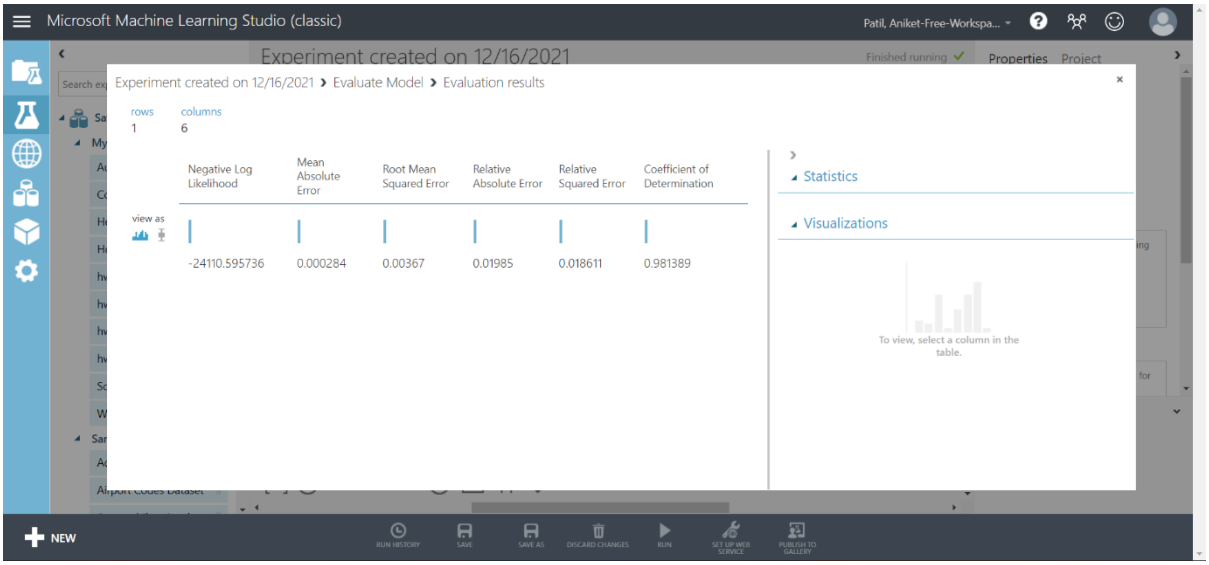
Number of random splits per node: 128, 132, 136

Minimum number of samples per leaf node: 1, 4, 5

10 Best Features:

1. Current Ratio
2. Working Capital to Total Assets
3. Quick Assets/Current Liability
4. Quick Ratio
5. Working capitcal Turnover Rate
6. Working Capital/Equity
7. Debt ratio %
8. Net worth/Assets
9. Equity to Liability
10. Liability to Equity

After training the model using only these 10 bests features the performance output is as follows:



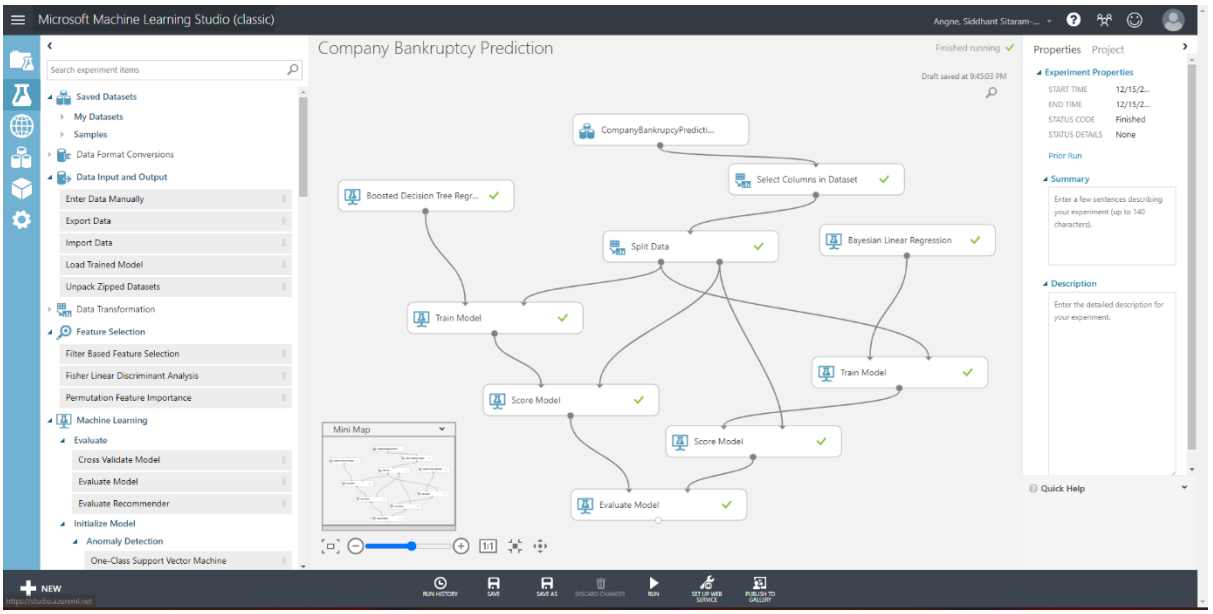
Model 2:

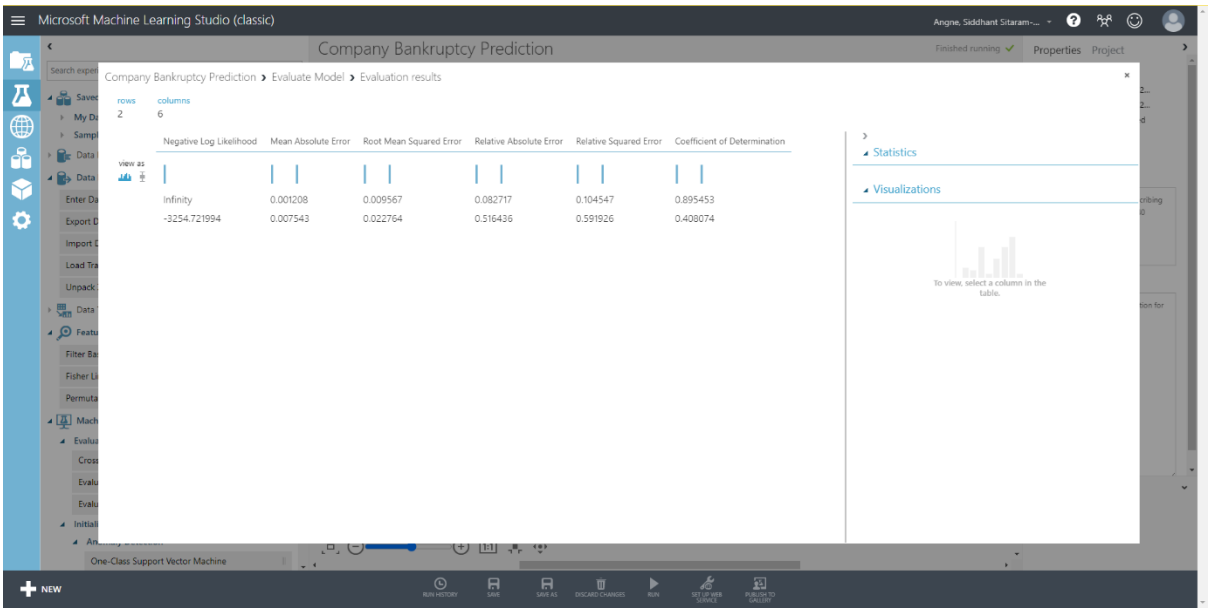
Member Name: Siddhant Angne

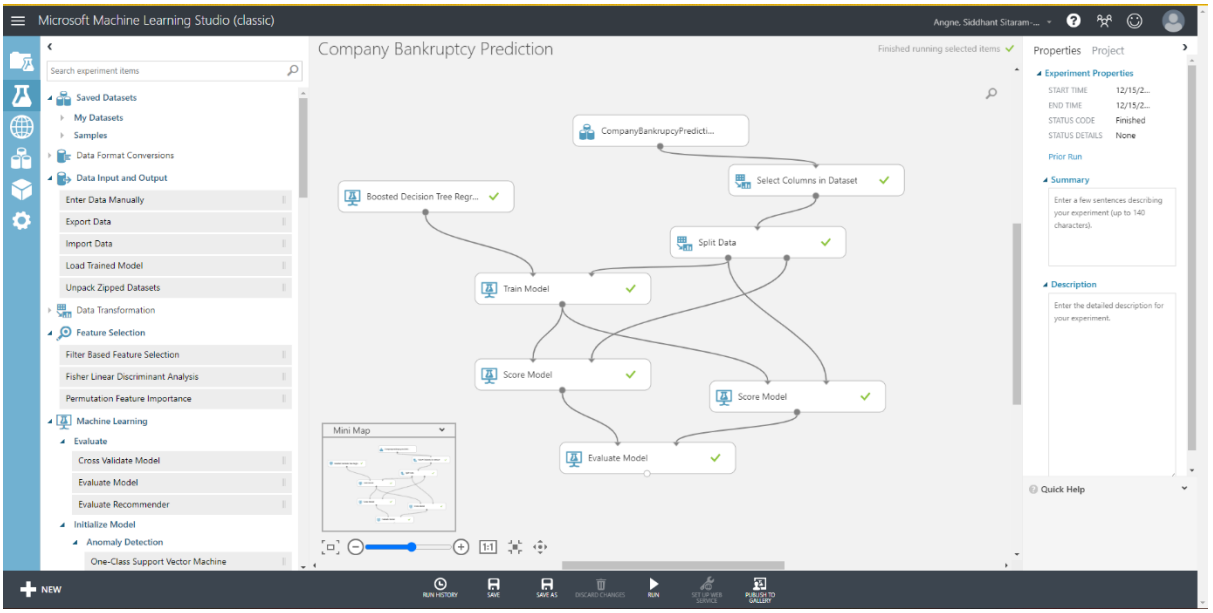
Algorithm Chosen for training the model: Since the problem is a regression problem, comparing Boosted Decision Tree Regression and Bayesian Linear Regression Algorithms to know which algorithm is the better suitable algorithm for this problem.

Pre-processing: Excluded Liability-Assets Flag, Net Income Flag columns from the dataset as these columns have only flag (0,1) data values.

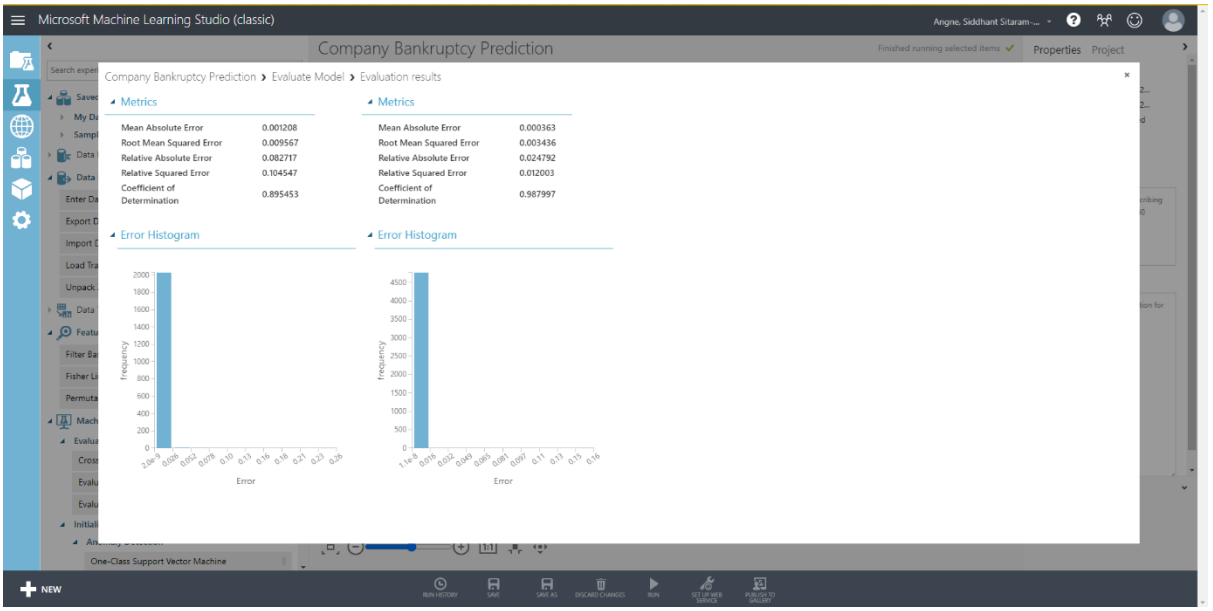
Best Algorithm: Boosted Decision Tree Regression Algorithm is the better algorithm as the Performance of Boosted Decision Tree Regression is 89% which is greater than the performance of Bayesian Linear Regression Algorithm which is 40%, hence the Boosted Decision Tree Regression Algorithm is chosen to perform the Company Bankruptcy Prediction.





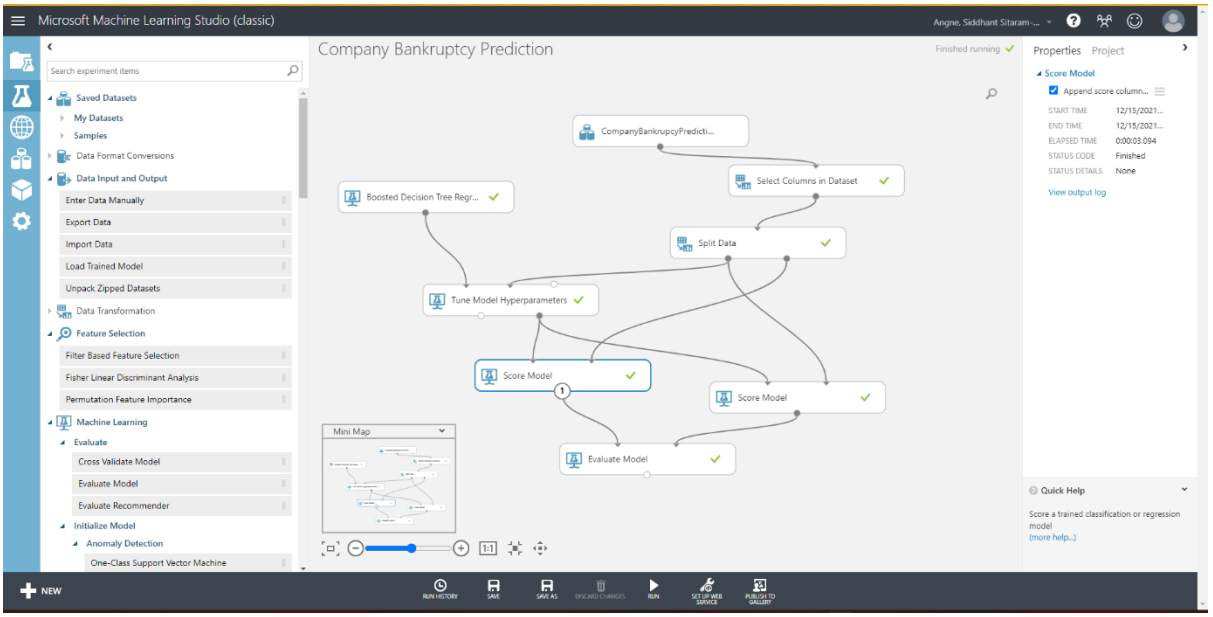


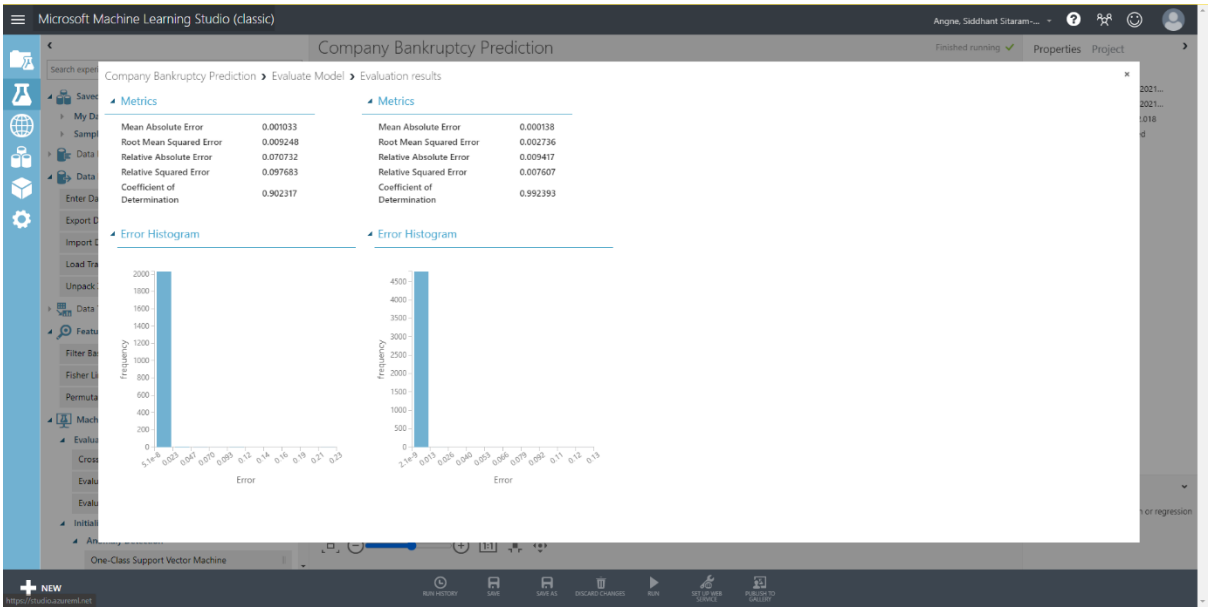
For the chosen algorithm, evaluate the degree of underfitting/overfitting problem by comparing the performance of the model for the scored training dataset with the performance of the model using the scored testing dataset. In the report, present model performance metrics and discuss underfitting/overfitting.

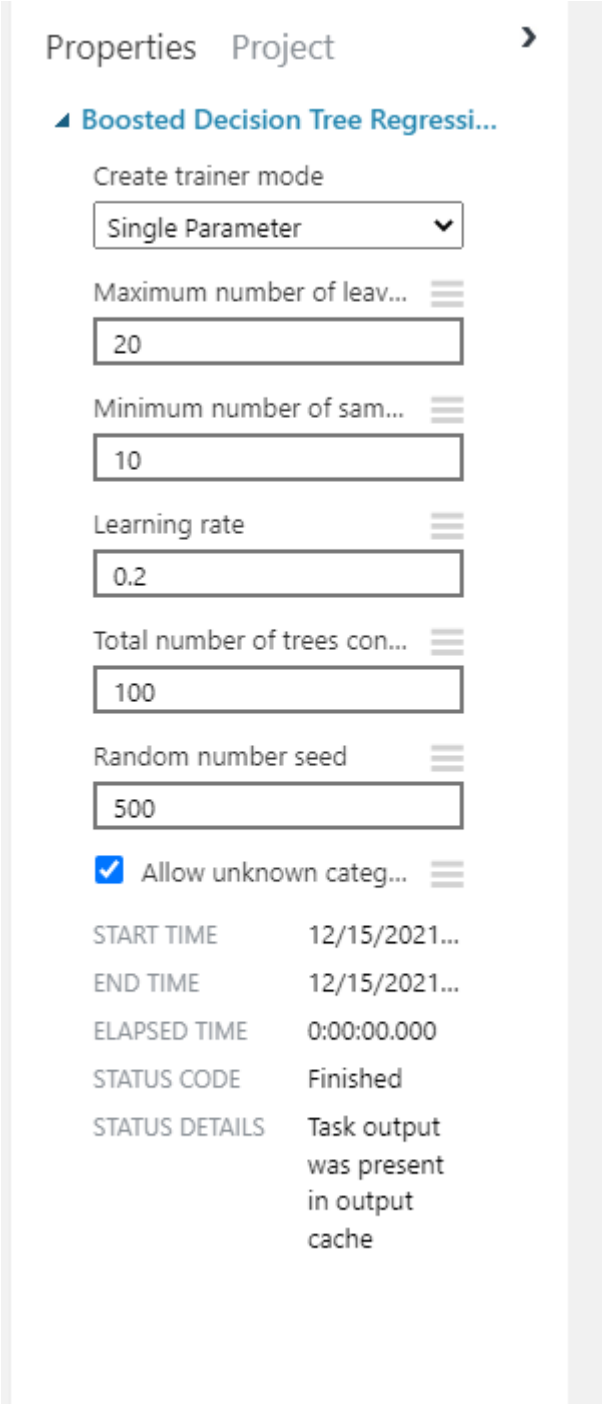


There is an Overfitting as the performance of the training dataset is higher than the performance of the testing dataset. The scored training dataset's performance is 89% and the scored training dataset's performance is 98%

After applying the Tune Model Hyperparameters the performance for the default parameters is 90% for the training dataset which improved sightly from the previous step.


Microsoft Machine Learning Studio (classic)
Company Bankruptcy Prediction
Finished running
Properties Project
Score Model
CompanyBankrupcyPredicti...
Select Columns in Dataset
Boosted Decision Tree Regr...
Split Data
Tune Model Hyperparameters
Score Model
Score Model
Evaluate Model
Mini Map
Quick Help
NEW


Microsoft Machine Learning Studio (classic)
Company Bankruptcy Prediction
Company Bankruptcy Prediction › Evaluate Model › Evaluation results
Metrics
Mean Absolute Error 0.001033
Root Mean Squared Error 0.009248
Relative Absolute Error 0.070732
Relative Squared Error 0.097683
Coefficient of Determination 0.902317
Metrics
Mean Absolute Error 0.000138
Root Mean Squared Error 0.002736
Relative Absolute Error 0.009417
Relative Squared Error 0.007607
Coefficient of Determination 0.992393
Error Histogram
Error Histogram
Frequency
Error
NEW



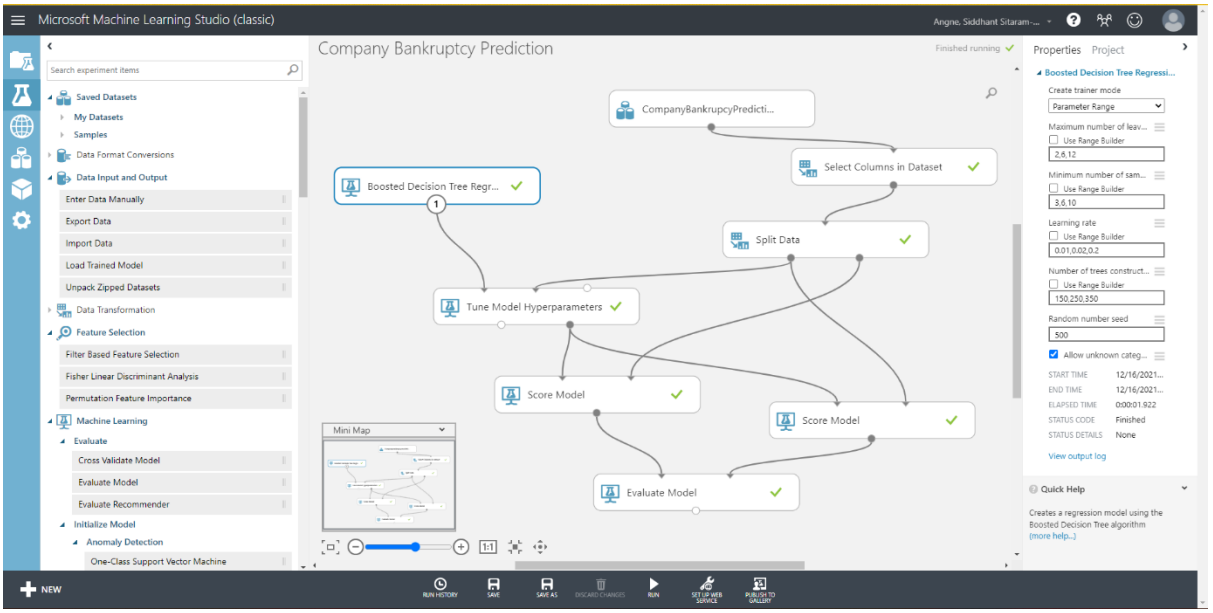
The default parameters are:

Maximum number of leaves per tree: 20

Minimum number of samples per leaf node: 10

Learning rate: 0.2

Number of trees constructed: 100



The parameter ranges are:

Maximum number of leaves per tree: 2,6,12

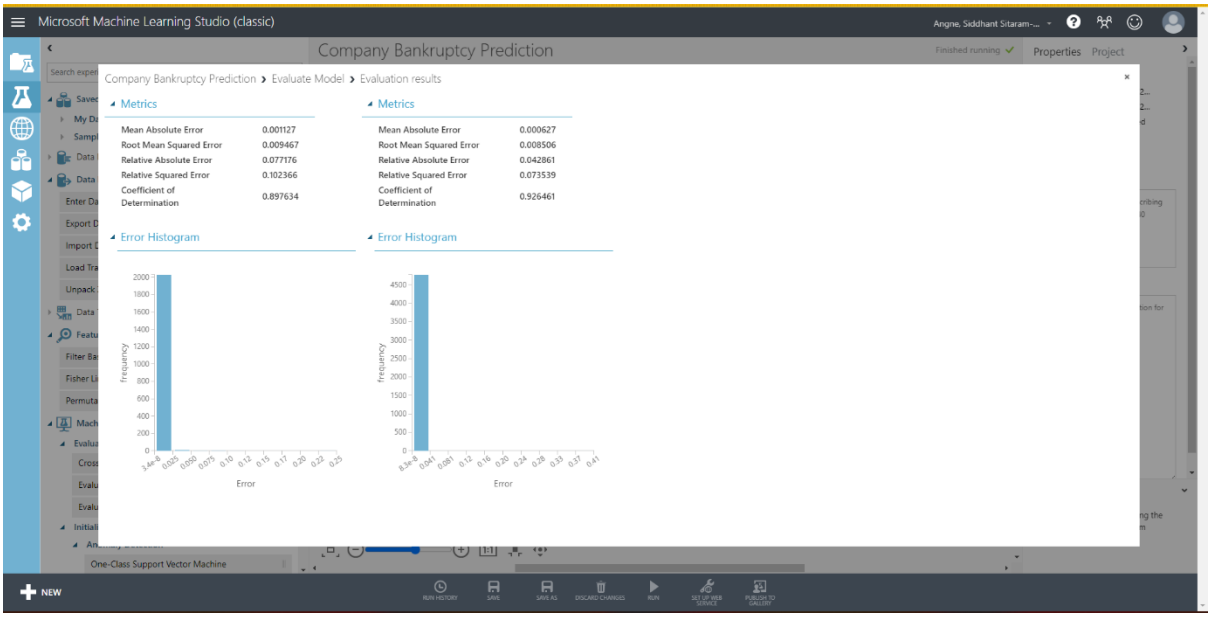
Minimum number of samples per leaf node: 3,6,10

Learning rate: 0.01,0.02,0.2

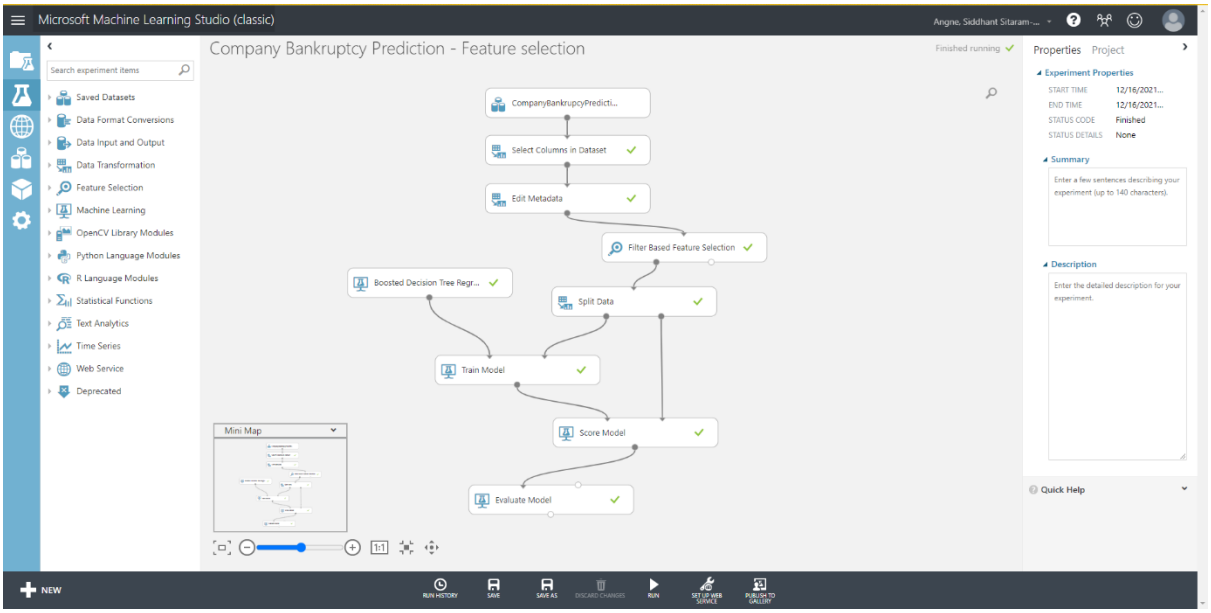
Number of trees constructed: 150,250,350

The performance of the best parameters is 89% for the training dataset.

The evaluate result of tune hyper parameters is as follows:



10 Best Features:



After applying the Filter Based Feature Selection module the top features are as follows

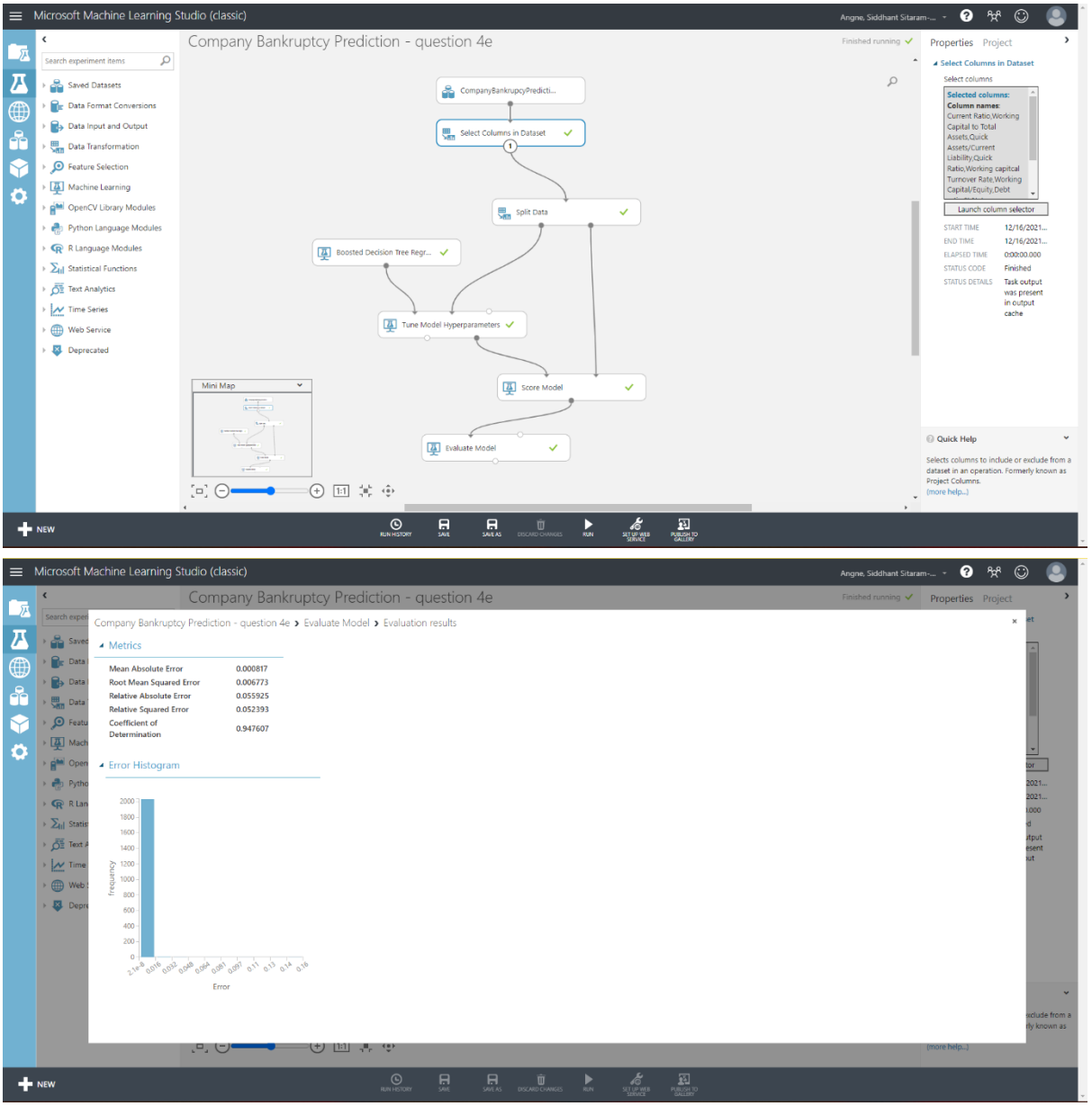
1. Current Ratio
2. Working Capital to Total Assets
3. Quick Assets/Current Liability
4. Quick Ratio
5. Working capitcal Turnover Rate
6. Working Capital/Equity
7. Debt ratio %
8. Net worth/Assets
9. Equity to Liability
10. Liability to Equity

Company Bankruptcy Prediction - Feature selection › Filter Based Feature Selection › Features

rows 1 columns 94

	Current Liability to Current Assets	Current Ratio	Working Capital to Total Assets	Quick Assets/Current Liability	Quick Ratio	Working capitcal Turnover Rate	Working Capital/Equity	Debt ratio %	Net worth/Assets	Equity to Liability	Liability to Equity
view as											
1	0.99912	0.890891	0.885847	0.867029	0.862279	0.714359	0.702851	0.702851	0.702847	0.69629	

After training the model using only these 10 bests features the performance output is as follows:



The performance after re-training the model with the 10 best features is 94%

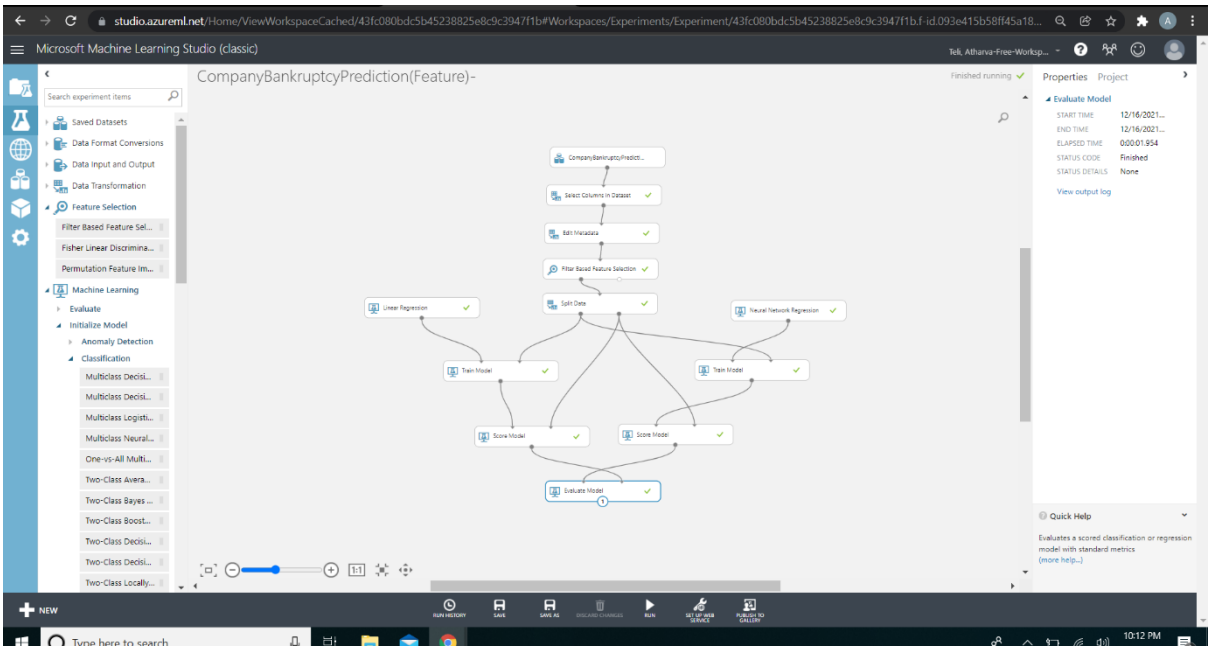
Model 3:

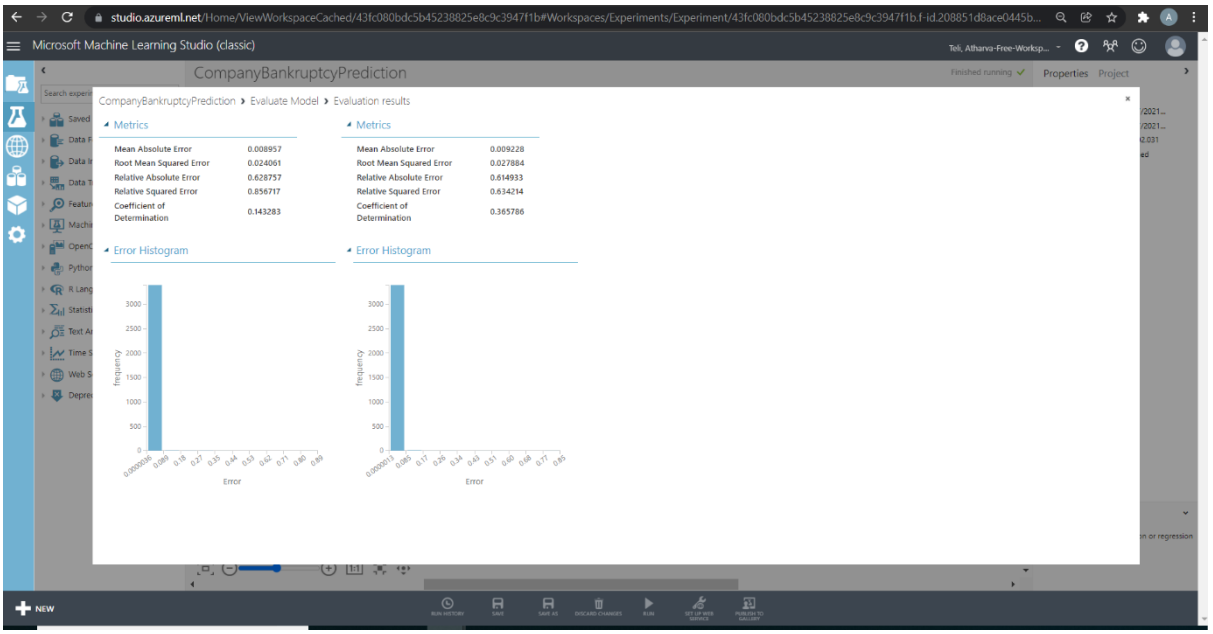
Member Name: Atharva Teli

Algorithm Chosen for training the model: Neural Network Regression and Linear Regression

Pre-processing: Excluded Liability-Assets Flag, Net Income Flag

Best Algorithm: Neural Network Regression

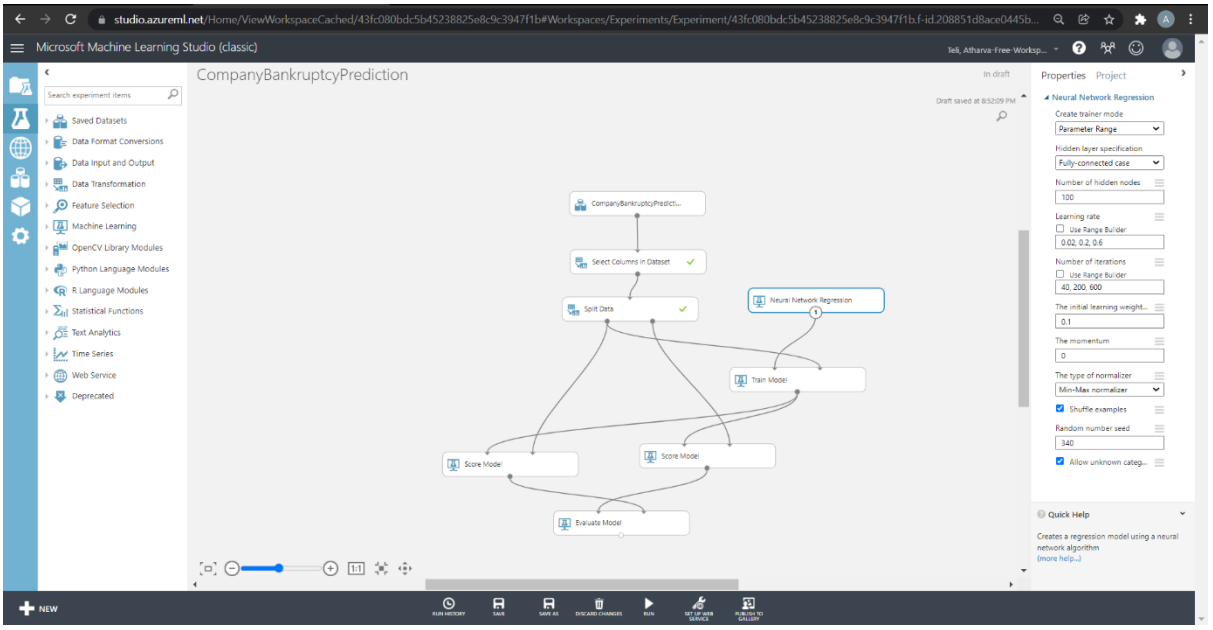


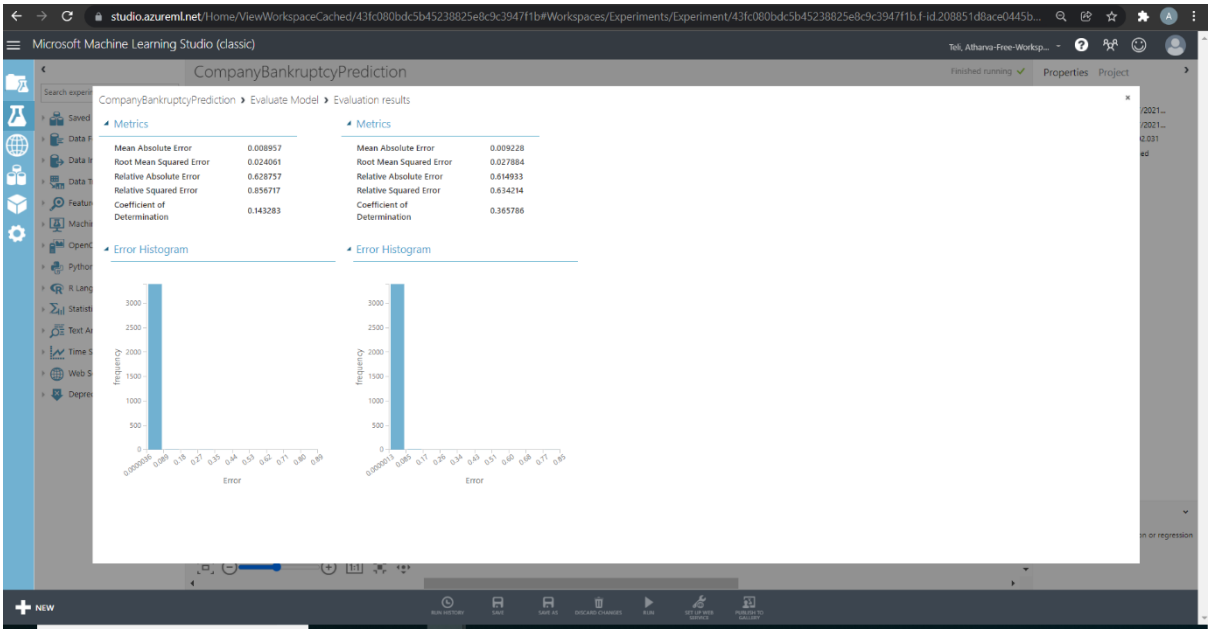


Performance of Neural Network Regression Algorithm is 48% which is greater than the performance of Linear Regression Algorithm which is 43%, hence Neural Network Regression

Algorithm is chosen to perform the Company Bankruptcy Prediction. Amongst all the regression modules Neural Network Regression Algorithm gave the best performance.

For the chosen algorithm, evaluate the degree of underfitting/overfitting problem by comparing the performance of the model for the scored training dataset with the performance of the model using the scored testing dataset. In the report, present model performance metrics and discuss underfitting/overfitting.





Since there is deviation between training dataset and testing dataset there is overfitting. The performance of the training dataset is 14% whereas the performance of testing dataset is 37%.

Default Tune Hyper Parameter Values:

Number of hidden nodes: 100

Learning Rate: 0.05

Number of iterations: 100

Initial learning weight: 0.1

The momentum: 0

Random Seed: 340

After applying the Tune Model Hyperparameters the performance went from overfitting to best fit and parameter range used as follows

Number of hidden nodes: 100

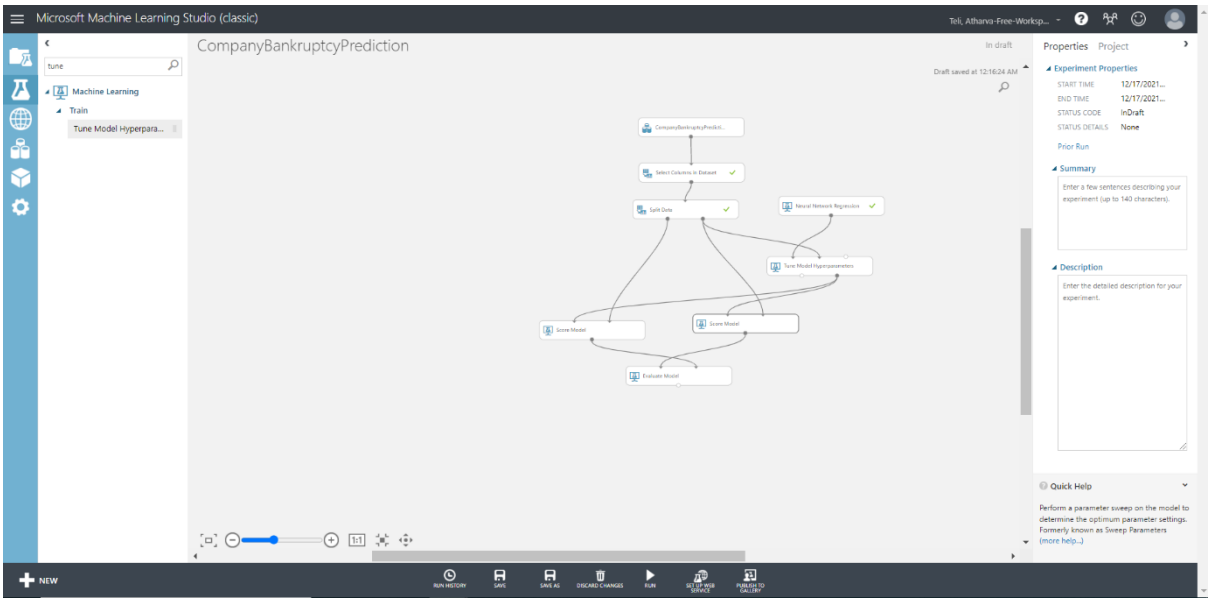
Learning Rate: 0.02, 0.2, 0.6

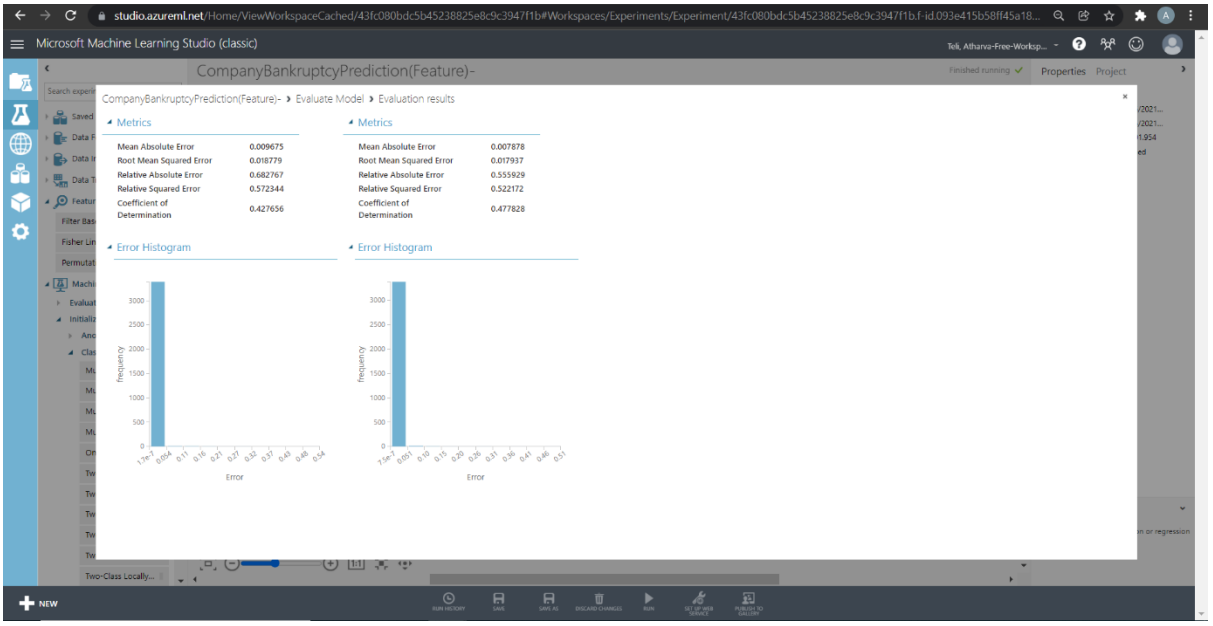
Number of iterations: 40, 200, 600

Initial learning weight: 0.1

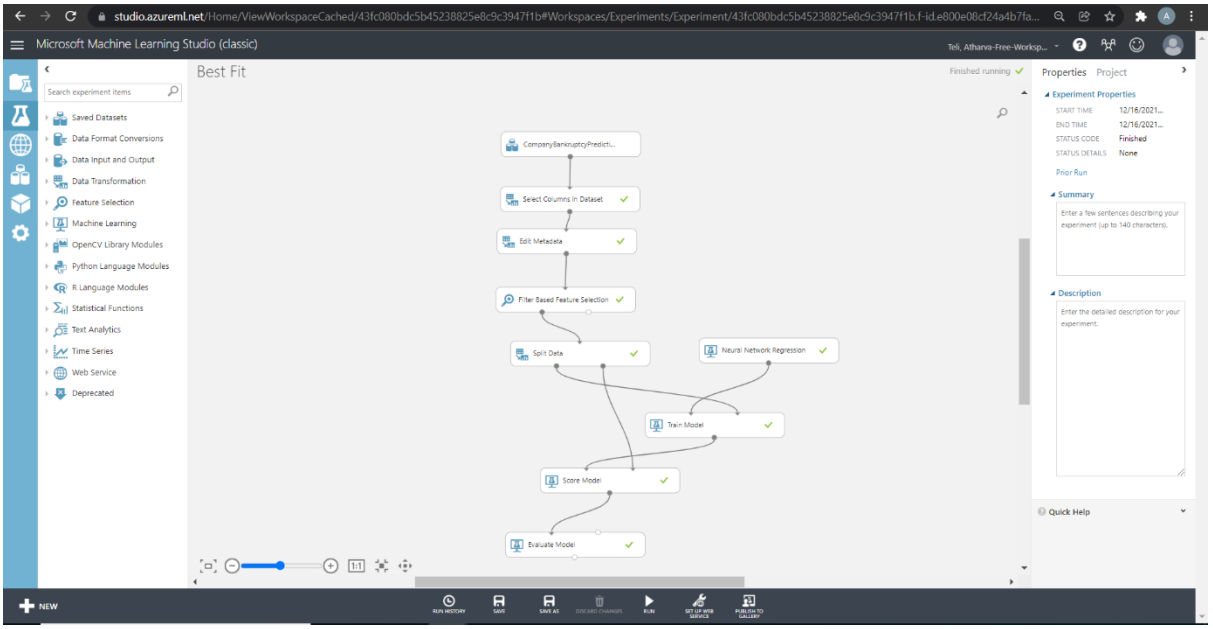
The momentum: 0

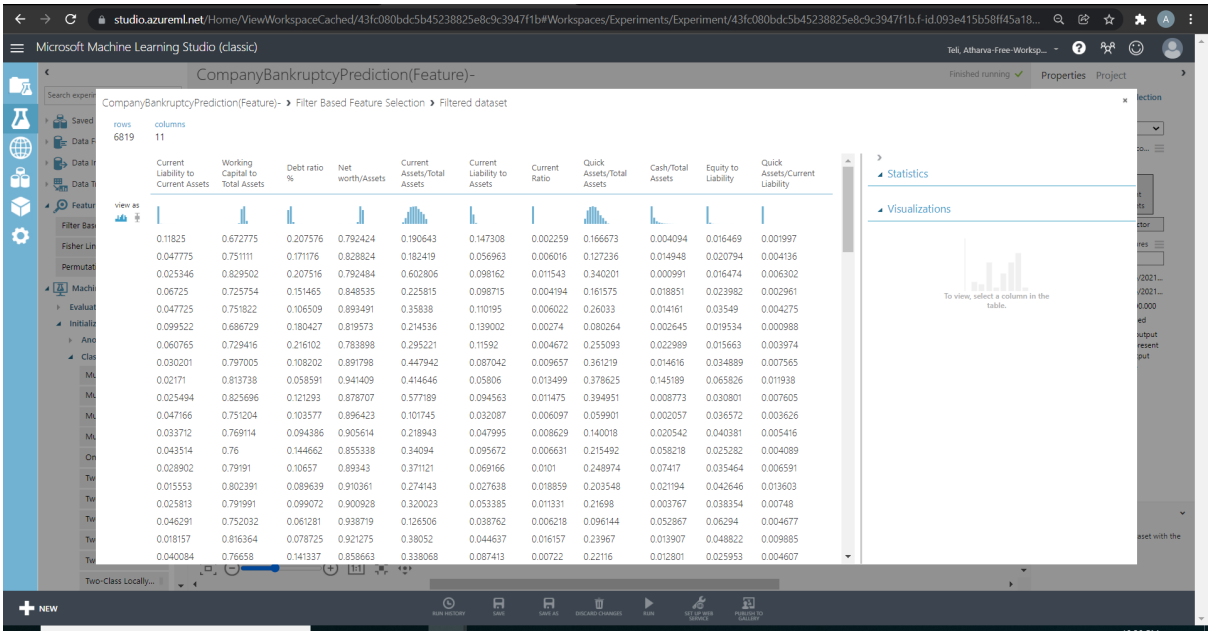
Random Seed: 340





The performance increased for Tune Model Hyper Parameter from 14% to 42%.



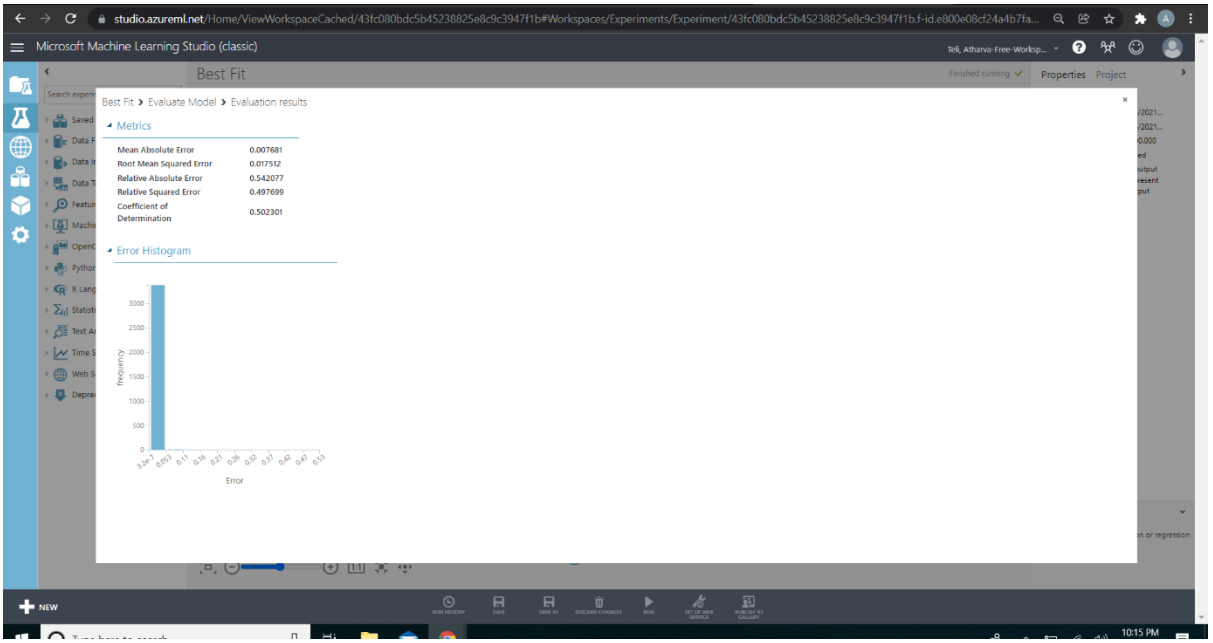


After applying the filter based feature selection module we got the top 10 features as follows:

10 Best Features:

1. Working Capital to Total Assets
2. Debt ratio %
3. Net worth/Assets
4. Current Assets/Total Assets
5. Current Liability To Assets
6. Current Ratio
7. Quick Assets/Total Assets
8. Cash/Total Assets
9. Equity to Liability
10. Quick Assets/Current Liability

After training the model using only these 10 bests features the performance output is as follows:



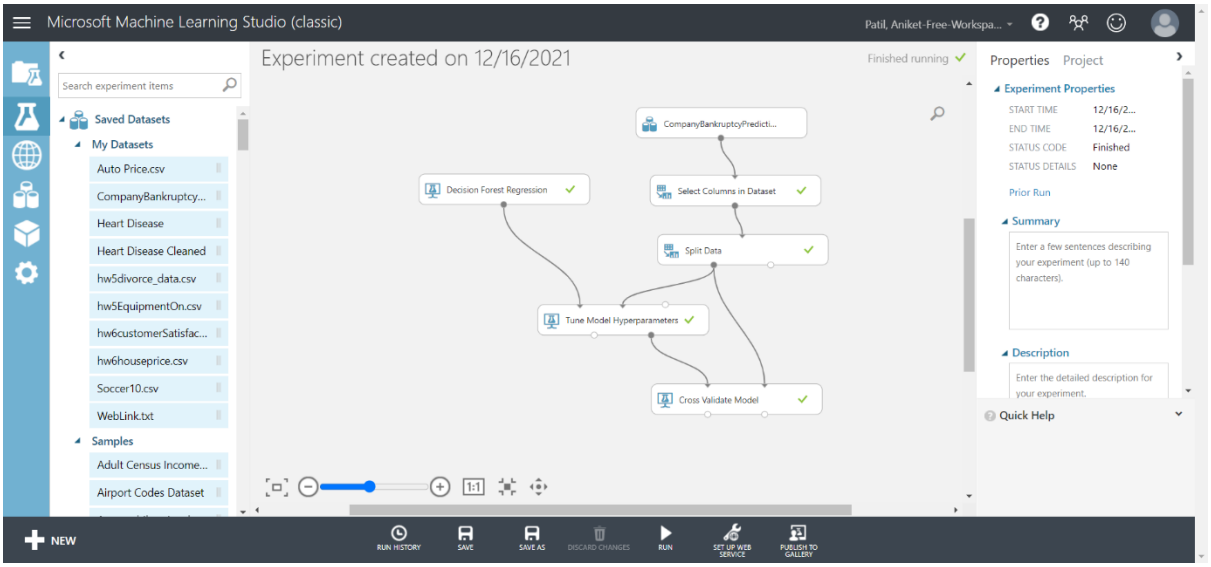
After retraining the model using only the best 10 features the result came out to be 50%.
For this Company Bankruptcy prediction problem, Neural Network Regression Algorithm is not the best option as the performance is low compare to other regression Algorithms.
Neural networks do not present an easily-understandable model. When looking at a decision tree, it is easy to see that some initial variable divides the data into two categories and then other variables split the resulting child groups. This information is very useful to the researcher who is trying to understand the underlying nature of the data being analysed.
Due to the limitation of the dataset, this algorithm is not as effective as compared to other algorithms.

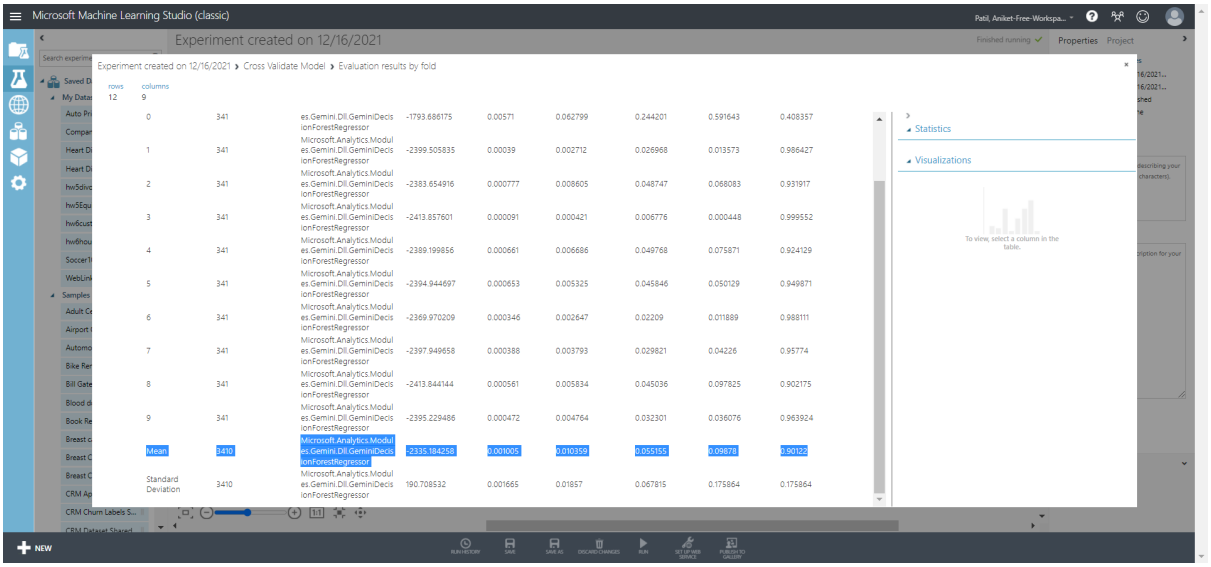
Finalizing the model:

Among all three-regression algorithm used for this problem, decision forest regression algorithm has the better performance with performance metric of 98% whereas the performance of the boosted decision tree regression algorithm is 94% which is second highest in performance and performance of Neural Network Regression Algorithm is 50% which is the third best performing model.

Therefore, decision forest regression algorithm was finalized.

Final ML model:





Conclusion:

The problem is predicting Bankruptcy for a company and as we identified it as a regression problem we used regression algorithms such as Neural Network Regression, Decision Forest Regression and Boosted Decision Tree Regression Algorithms to find out which Algorithm predicted the best performance metrics for the given problem. Each regression algorithm provided with various outputs resulting in overfitting, underfitting and best fitting performance metrics. Then we used the tune model hyperparameters module to fine tune the model to get

best performance for training the model. After that we applied these setting along with the filter based selection module to get the performance using the10 best features and compared each algorithm performances with each other. After comparing the results of the regression algorithms we were able to finalize with the Decision Forest Regression Algorithm providing with the best performance of 98%.
If this was a real life project, the follow-up steps would be:-