DS 5220: Supervised Machine Learning

Project:
Bankruptcy Prediction

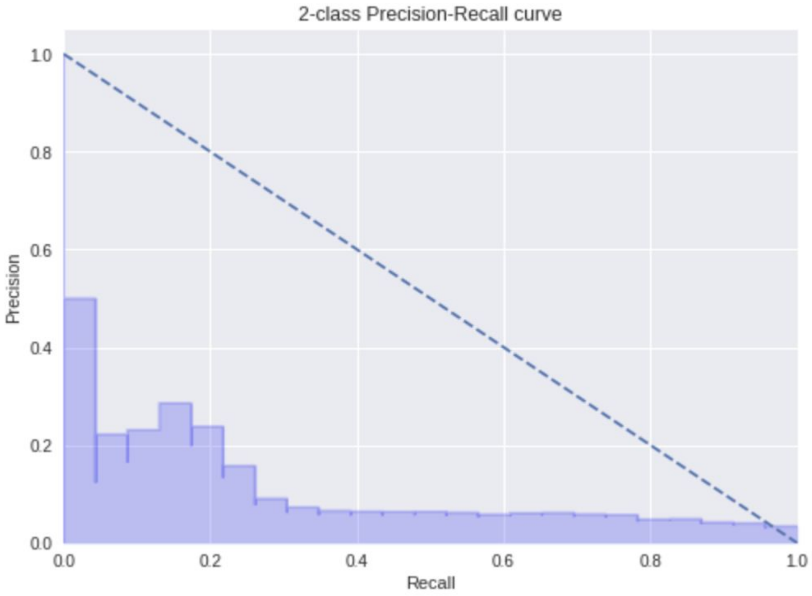
Group 2:
Ayush Bhandari
Camellia Debnath
Viral Pandey

The Project So far

We completed the following steps by our Project Pitch:

- Exploratory Data Analysis
- Data Imbalance Issues
- Missing Data Issues
- Baseline Model: Logistic Regression

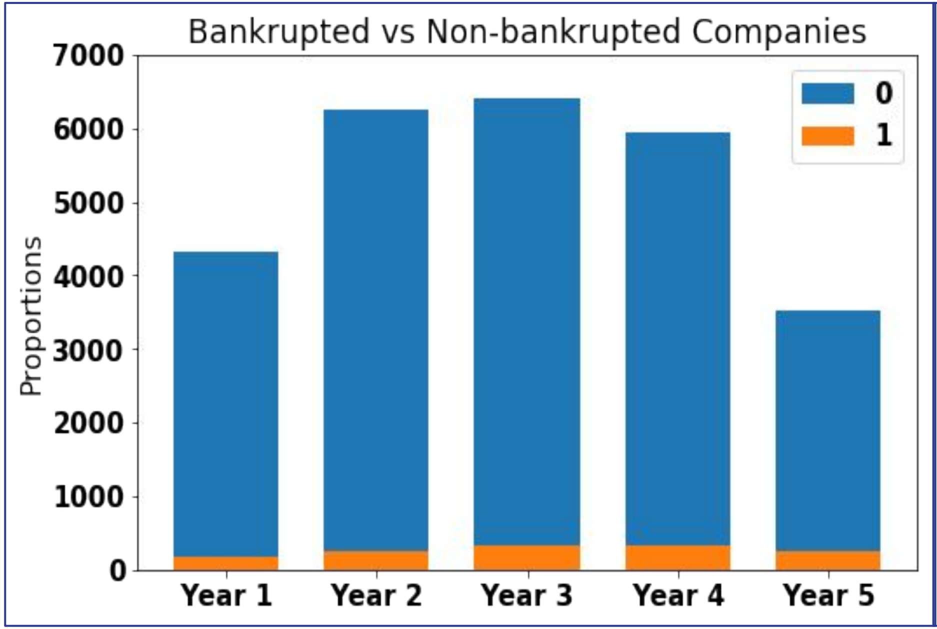
- ❏ Specificity: 0.99412
- ❏ Sensitivity: 0.043478
- ❏ F1 score: 0.942710



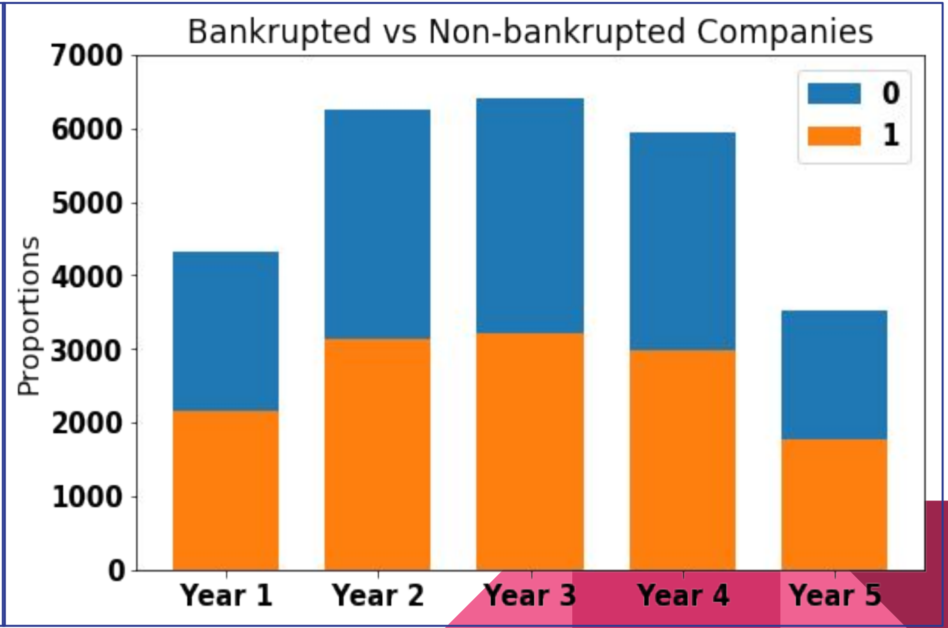
Dataset Characteristics

- Data Imbalance: SMOTE (with sampling strategy)

Before oversampling

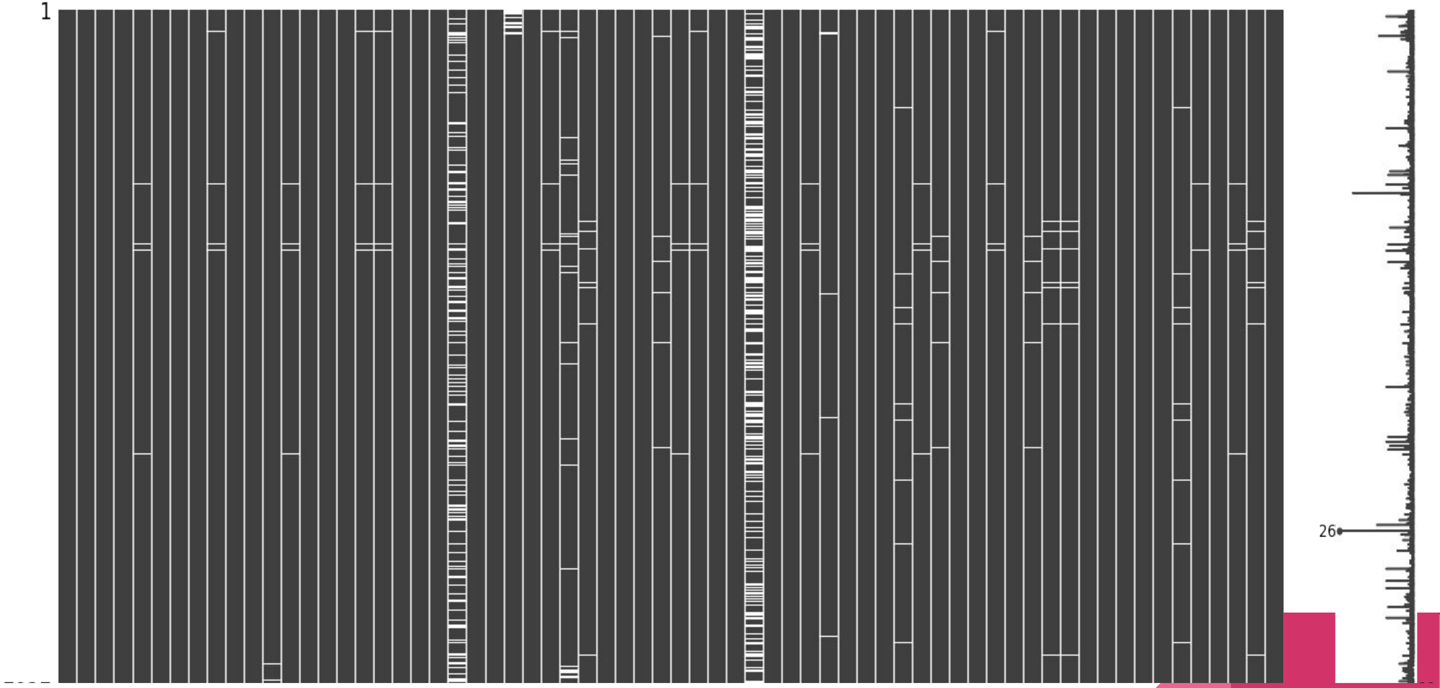


After oversampling



Dataset Characteristics contd.

- Imputation (Mean)



X axis: Attributes 1 through 64

Classification Models

- Logistic Regression
- Naive Bayes
- Nearest Shrunken Centroids
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Support Vector Machines
- Random Forest
- XGBoost
- Neural Networks

Why fit so many models?

- Academic Curiosity
  - *Effects of data imbalance, dimensions*

- Performance Comparison
  - *Accuracy, Precision, Recall, Ease of fit*

Logistic Regression

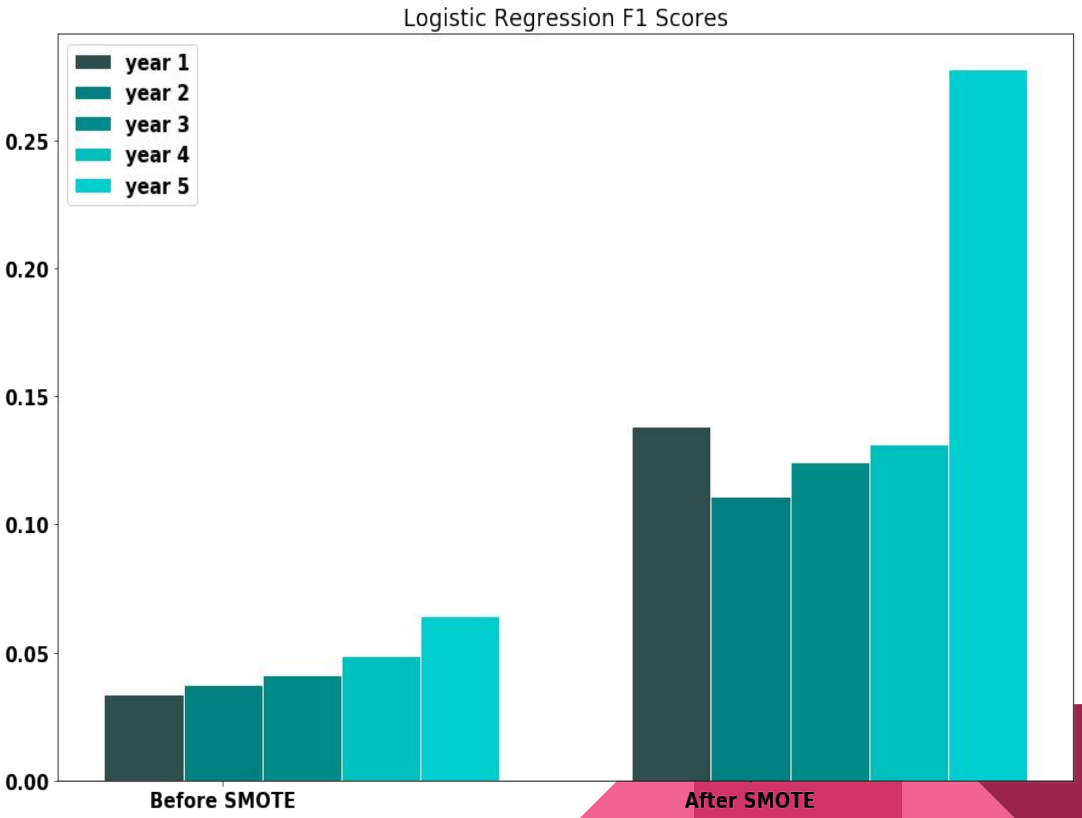
We fit Logistic Regression models before and after oversampling the minority class using SMOTE. We did not notice much improvement.

Classification report for Year 5

	precision	recall	f1-score	support
0.0	0.94	1.00	0.97	1106
1.0	0.00	0.00	0.00	76
micro avg	0.94	0.94	0.94	1182
macro avg	0.47	0.50	0.48	1182
weighted avg	0.88	0.94	0.90	1182

Classification report for Year 5

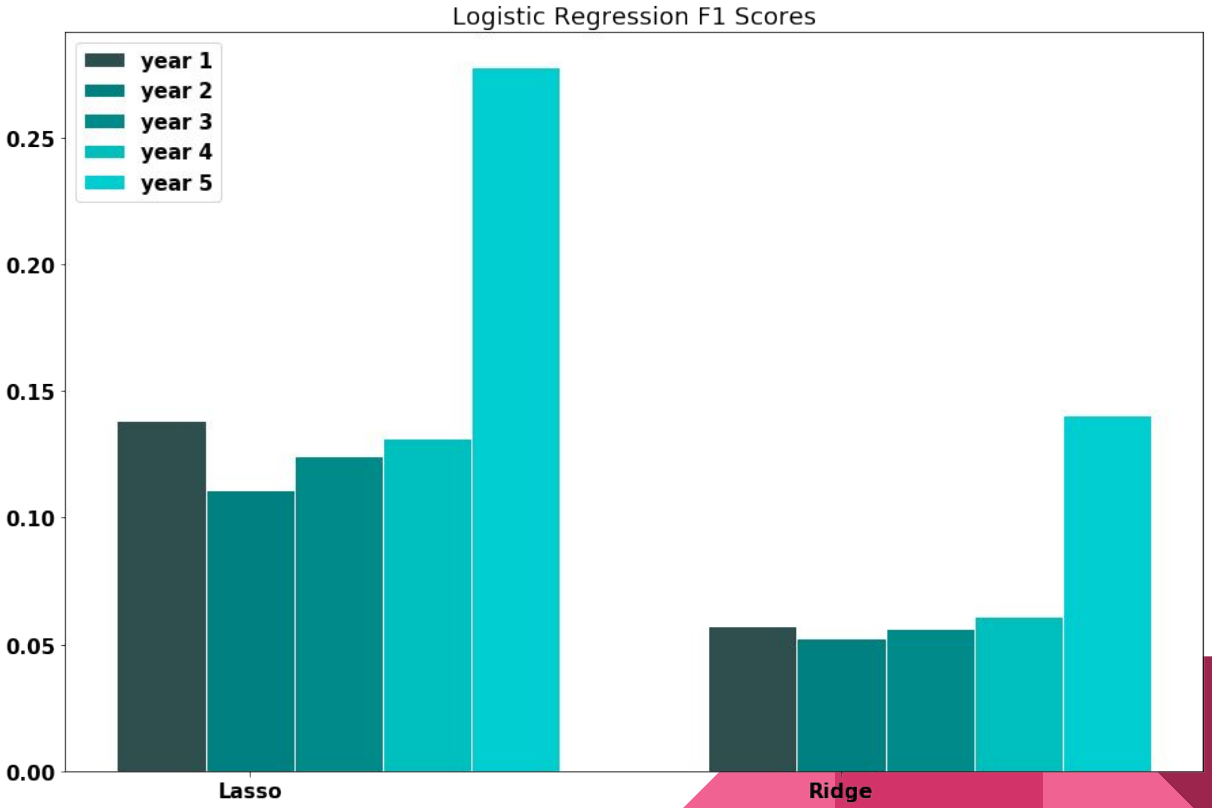
	precision	recall	f1-score	support
0.0	0.97	0.79	0.87	1106
1.0	0.18	0.66	0.28	76
micro avg	0.78	0.78	0.78	1182
macro avg	0.57	0.72	0.57	1182
weighted avg	0.92	0.78	0.83	1182

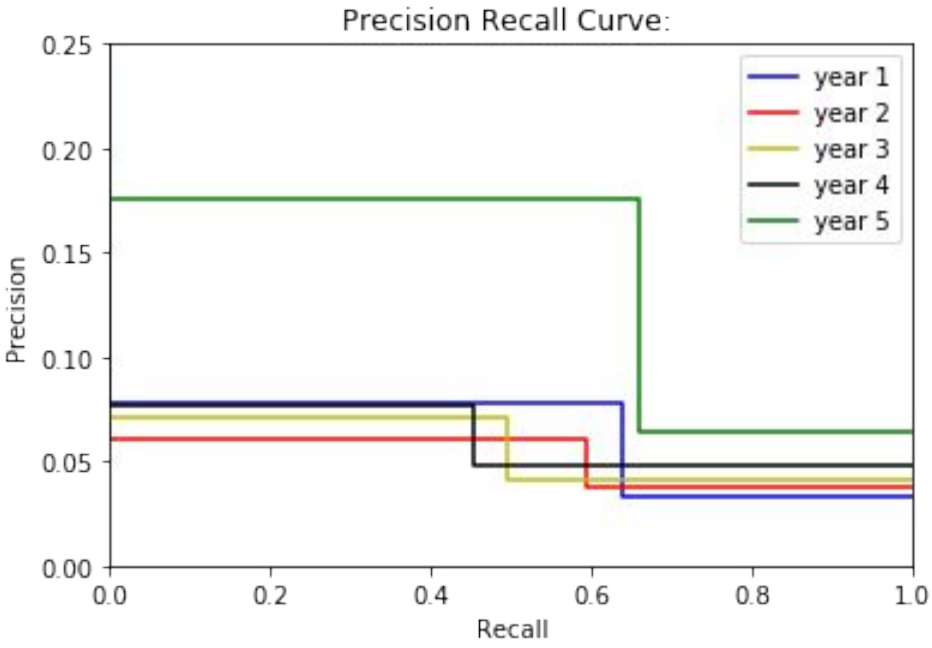


Logistic Regression

We fit Logistic Regression models using different regularization methods like Lasso and Ridge. Lasso seems to perform a little bit better than Ridge.

Possible reason: Since we did not do any kind of feature selection, Lasso driving the coefficients of insignificant variables to zero helps build a better model.





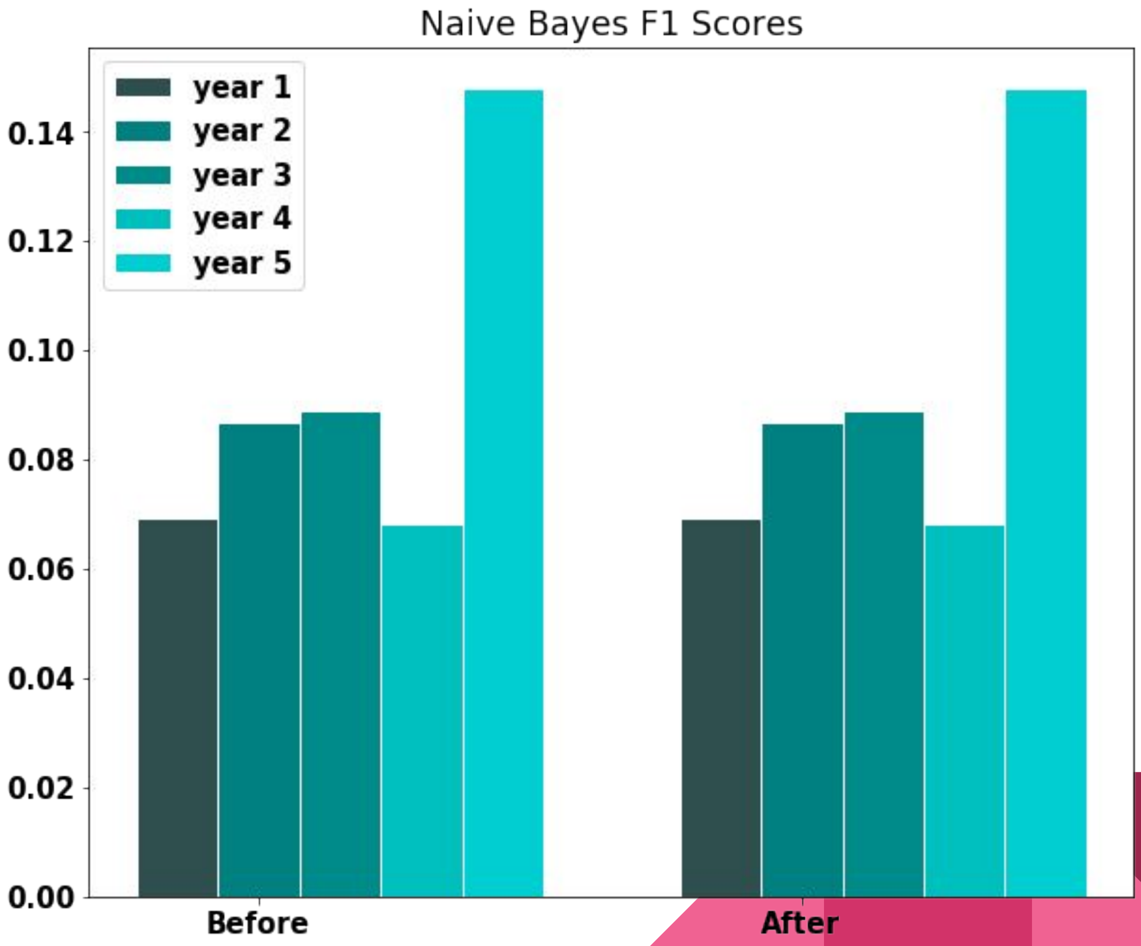
PR Curve

For Logistic Regression

Naive Bayes

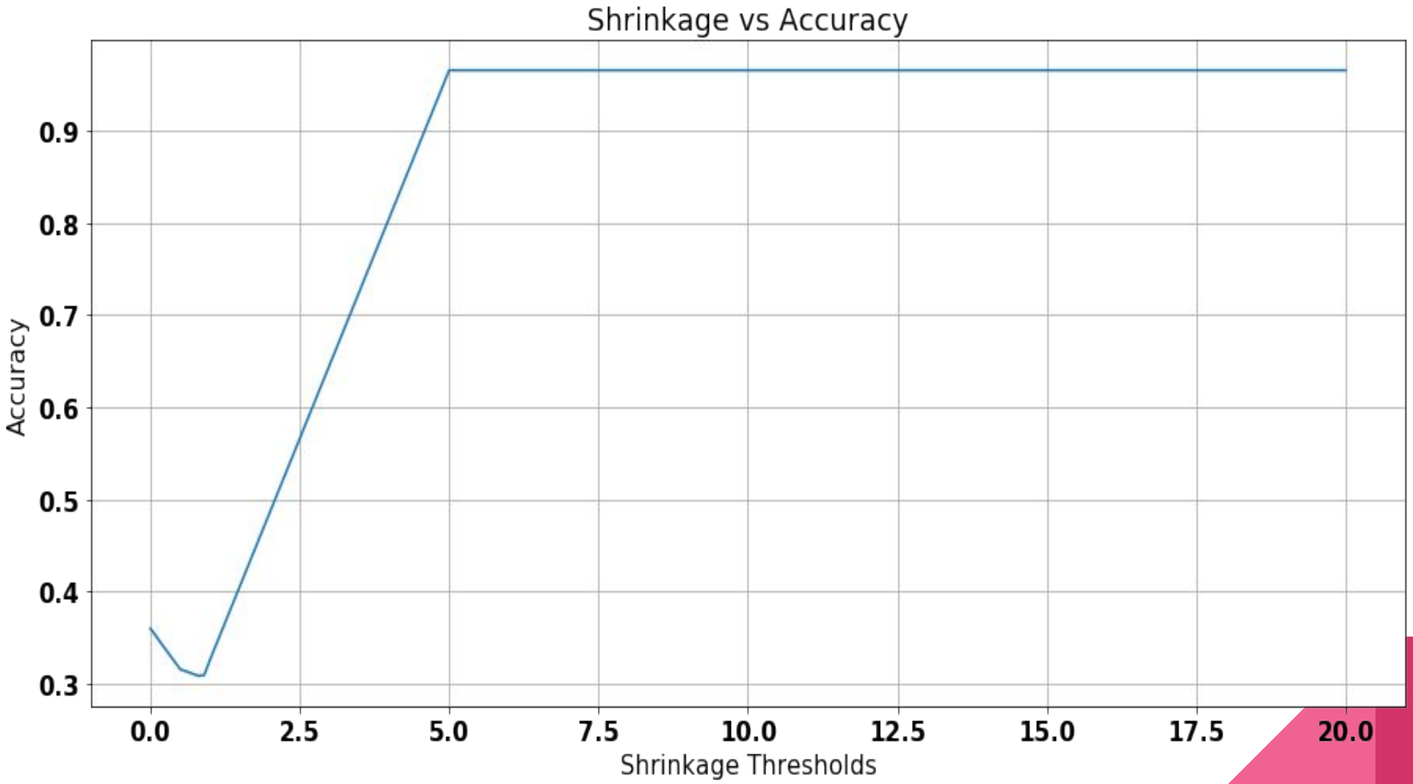
We fit Naive Bayes models before and after removing the correlated features. We did not notice any improvement.

Possible reason: Uncorrelated, but dependent features.



Nearest Shrunken Centroids

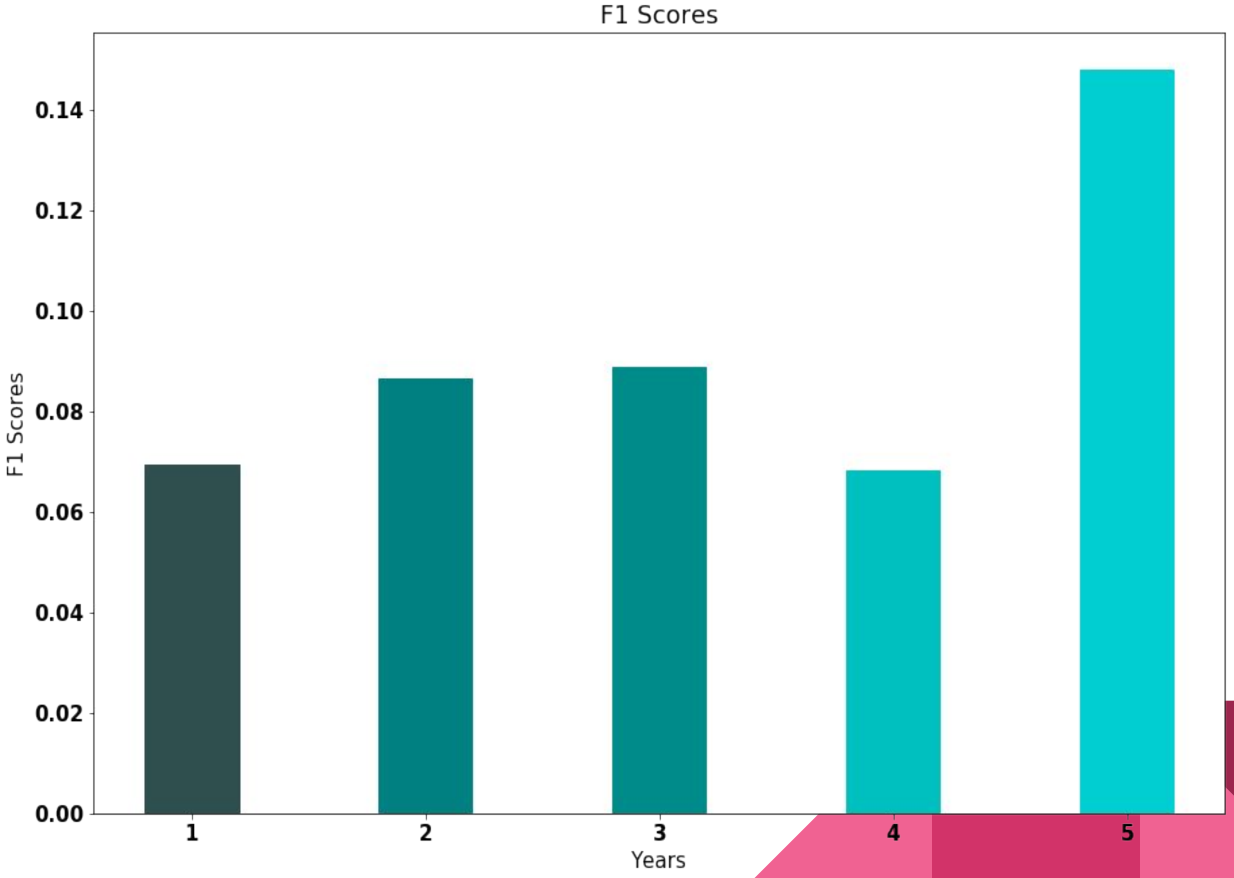
- Cross Validation to Find best Shrinkage Factor



Nearest Shrunken Centroids contd.

Shrink Threshold = 5 (From cross Validation)

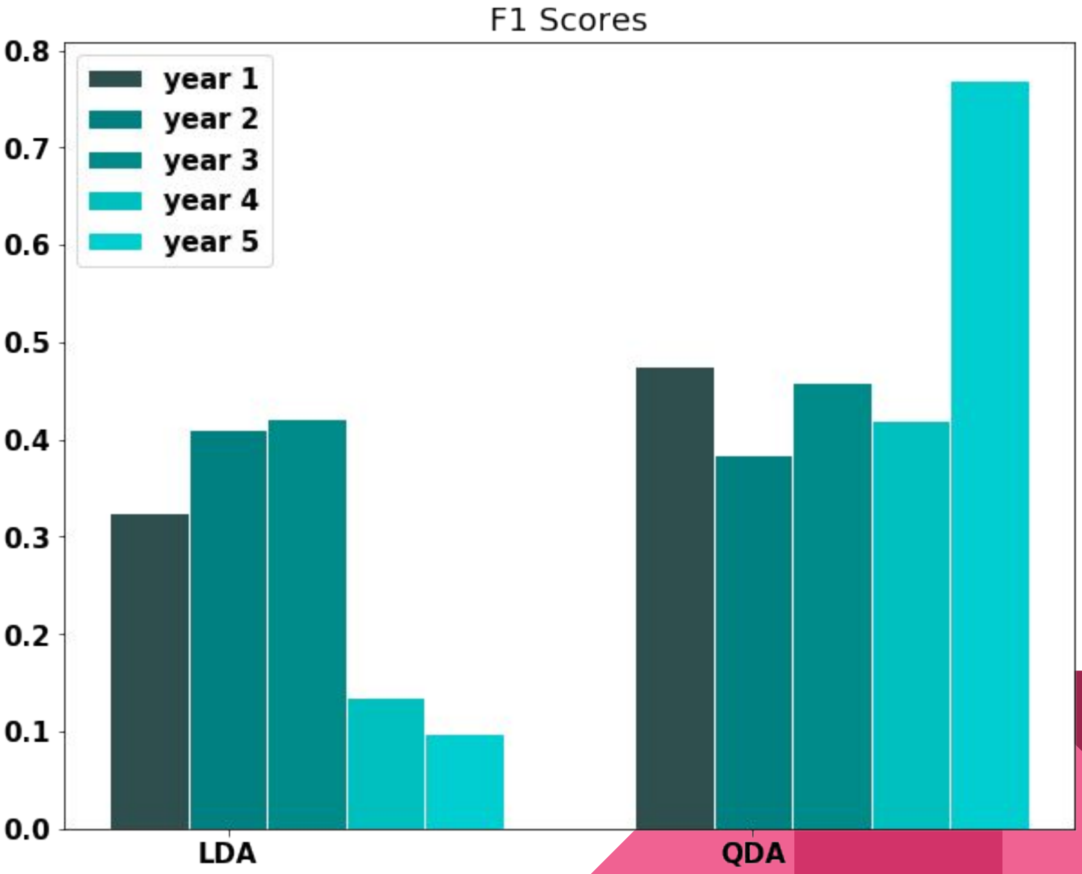
Again Year 5 has the best performance



Linear And Quadratic Discriminant Analysis

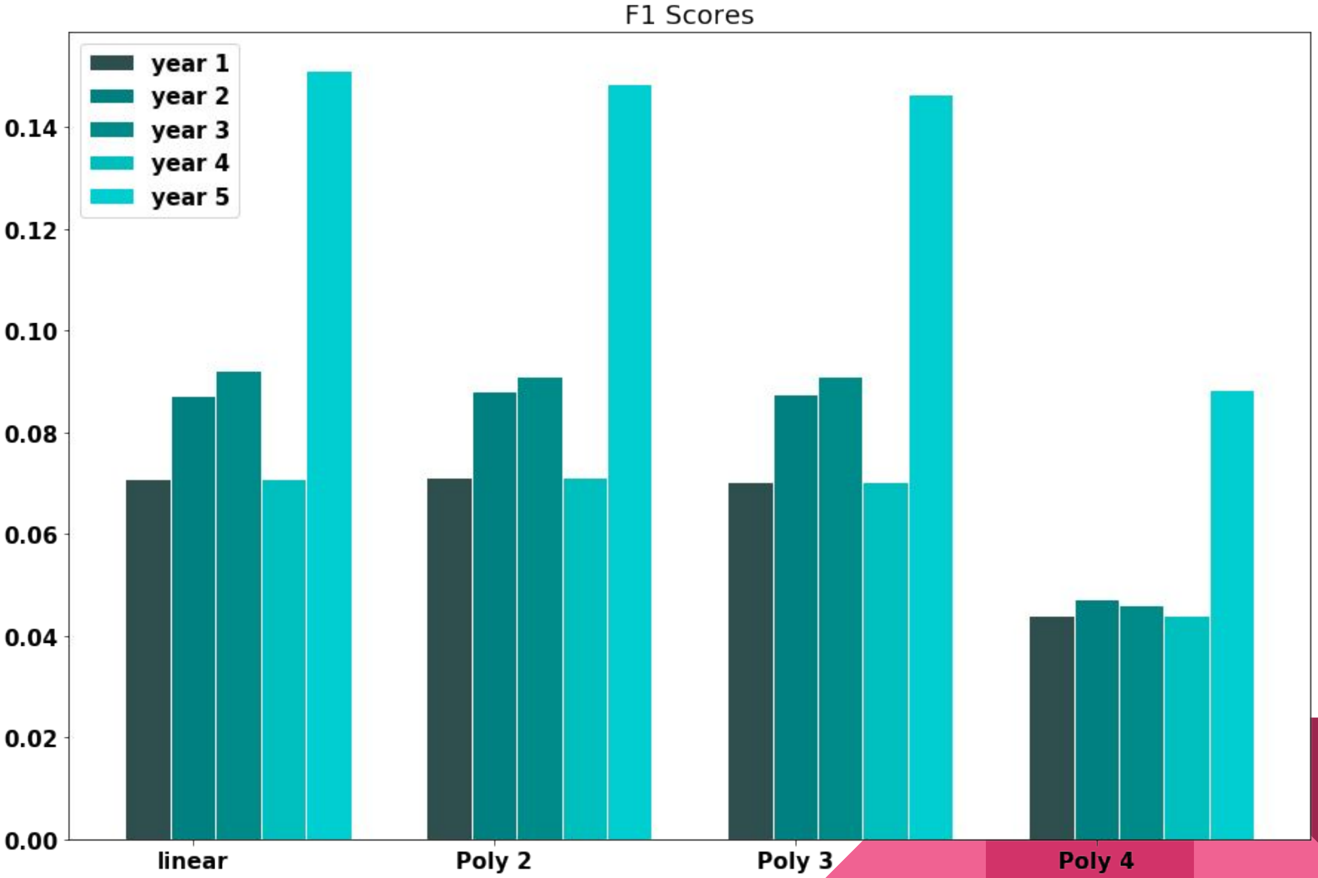
QDA performs much better than LDA, especially for year 5 dataset.

Possible reason: It’s unreasonable to assume that all the predictors have the same covariance matrix. QDA arises from Discriminant Analysis when we forego that assumption.



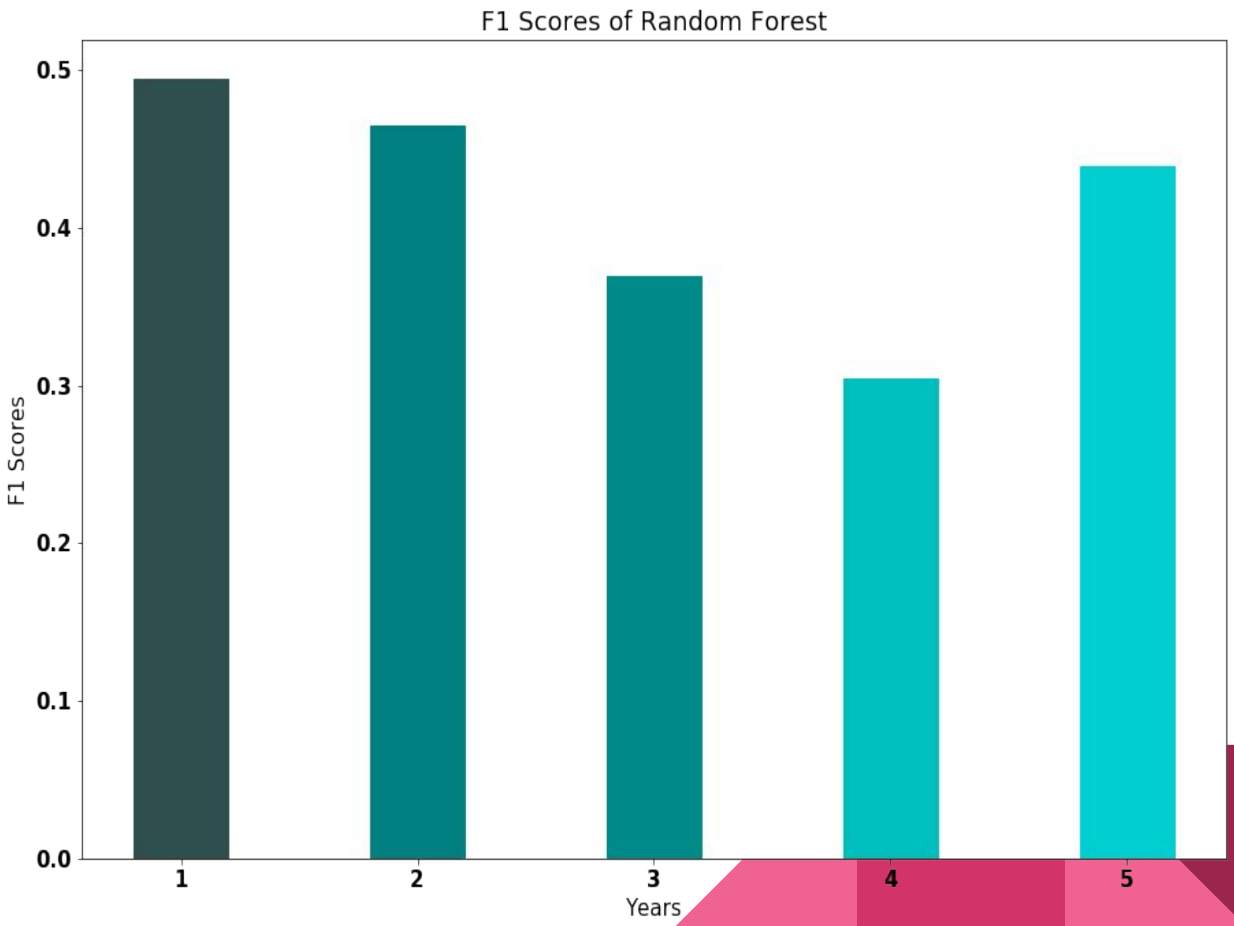
Support Vector Machines

SVM does not perform well for this dataset. Also, we notice that the performance drops when we increase degree of the kernel polynomial.



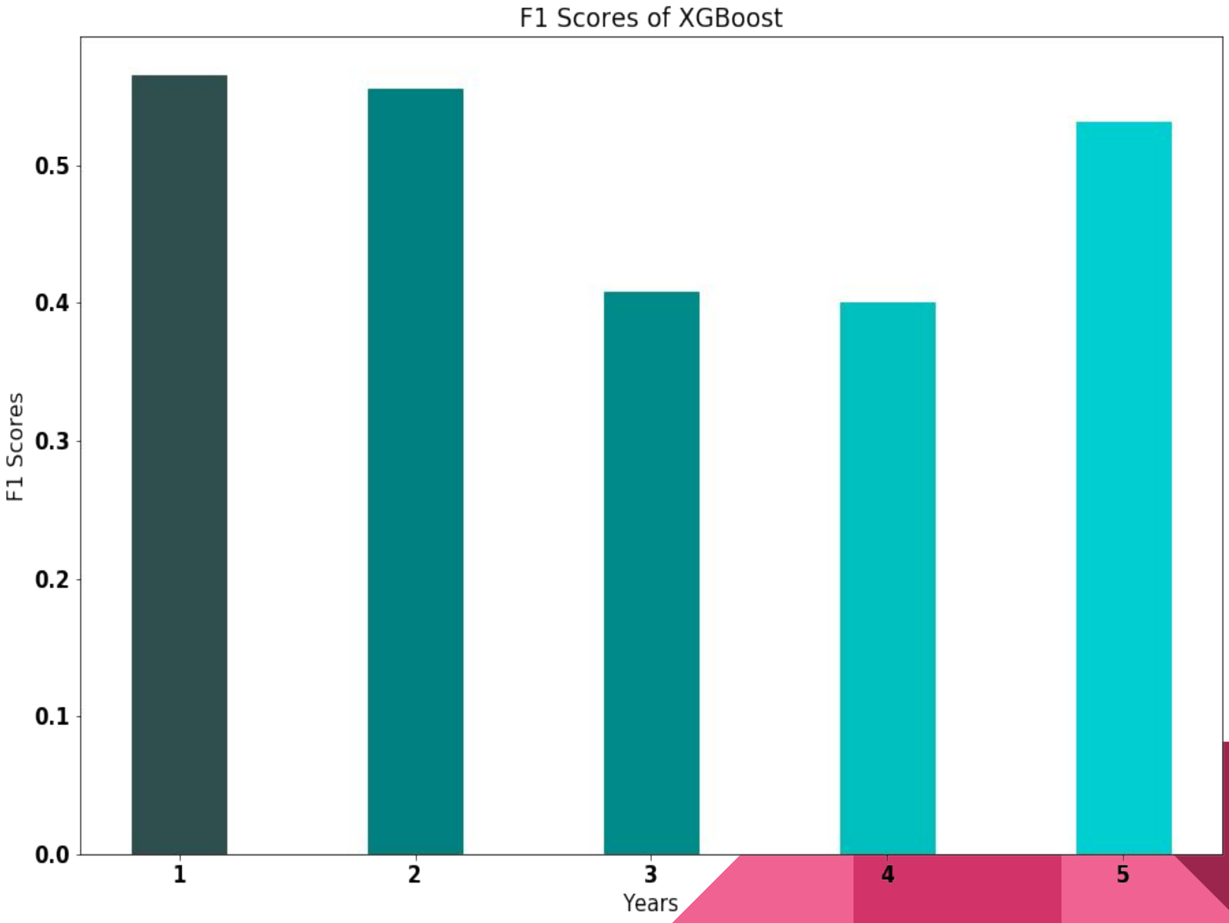
Random Forests

Random Forests perform better than SVM but less than QDA. The number of estimators were 200 with maximum depth of 25.



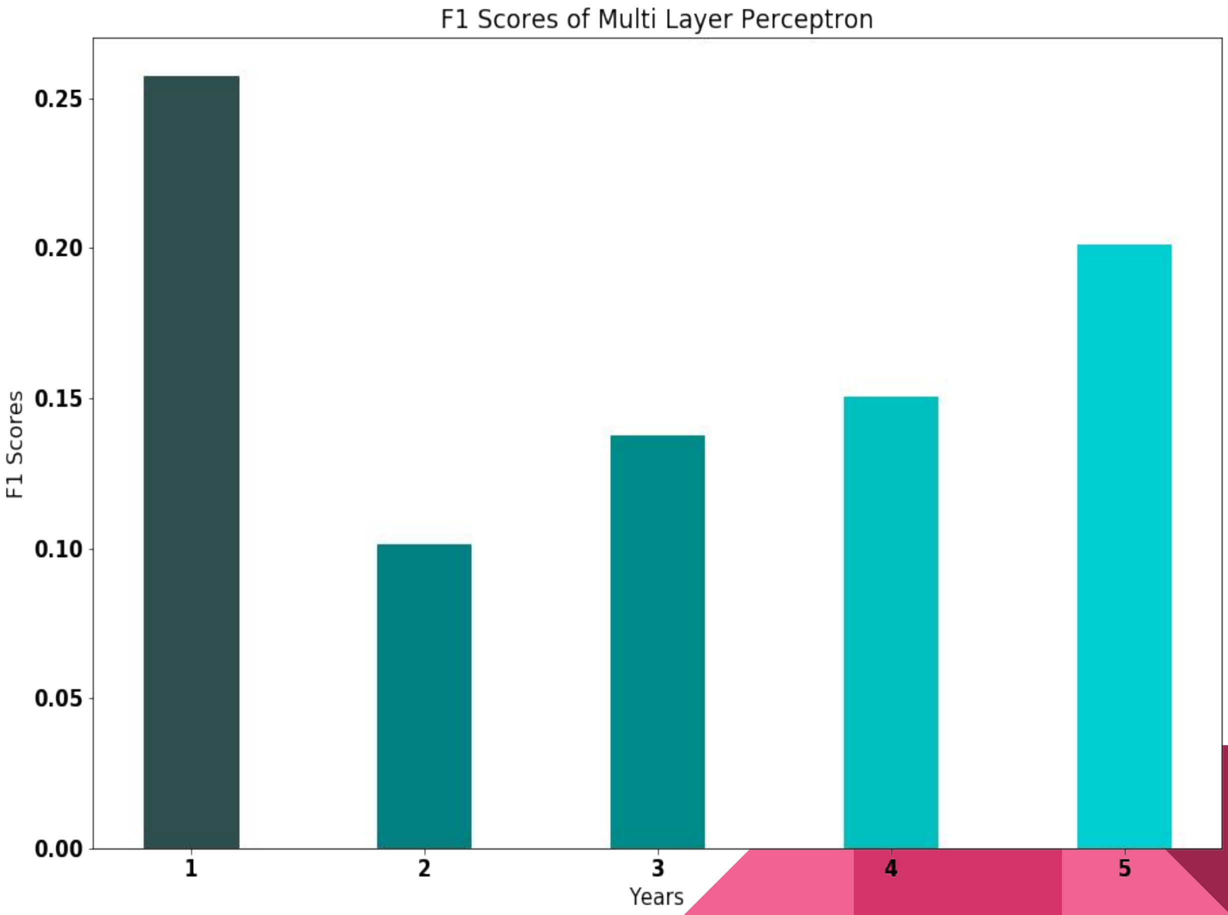
XGBoost

XGBoost performs better than Random Forests as expected. The maximum depth here was 10 and lambda was 0.5.



Neural Networks

Surprisingly, neural network failed to perform well. The network had 1 hidden layer with 10 neurons, learning rate of 0.01 and adam optimizer using adaptive gradient descent.



Inferences

- Domain Knowledge:
    - Correlated Features: most of the predictors were ratios related to profit/sales, eg gross-profit/sales, net-profit/sales, profit+depreciation/sales etc.
    - High Coefficients for features related to profit and liabilities
- Machine Learning Perspective:
    - Performance of Classification Methods
        - Data is not linearly separable, Linear classification methods do not work well
        - Non-linear Methods
        - Dependency Issues
        - Hyperparameter Tuning
        - Best Models

Questions?