Introduction

Benefits of Bankruptcy Prediction

Early warning, Risk management, key variables , major issue and Increased Transparency

Process of Analysing the data

We subsetted and resampled our datausing bootstrap to build our model using 3 train sets and performed the prediction on the test data set using logistic reg and decision tree to see which one provides the best model

Identifying Significant Variables

Identifying which financial ratios are highly significant

Recomendations for companies

Monitor performance and pro-active decision making

Dataset and EDA

Our dataset is taken from Kaggle's website, contains around 6819 obv with 96 variables.

Generic Logistic Regression

We reduced our 95 financial variables down to 70 using domain knowledge and ran a logistic regression to see which variables are highly significant

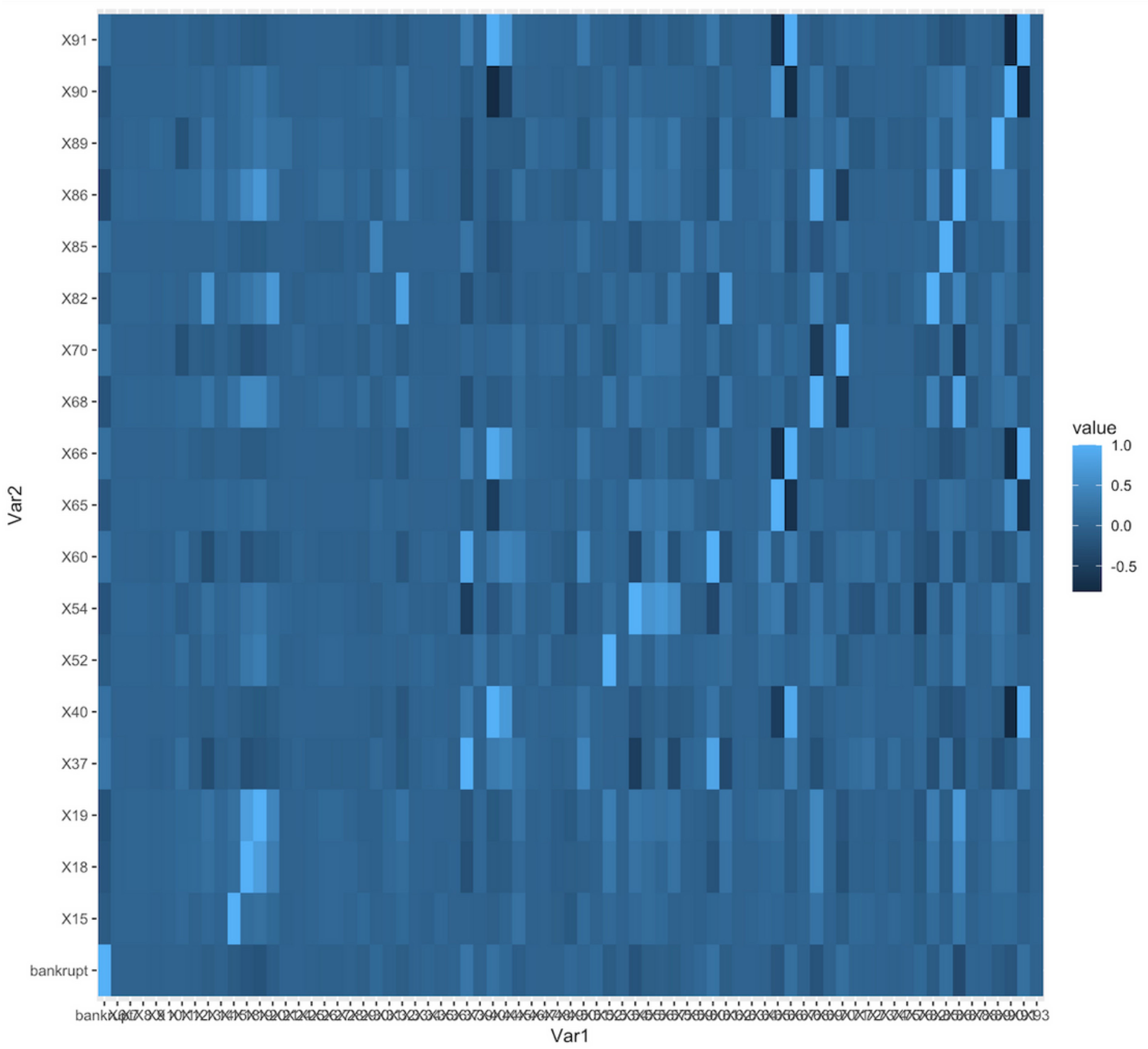
Collinearity Reduction

We further reduced our 70 financial variables down to 18 variables by checking the collinearity among the variables that are highly correlated to bankrupt

18 Financial Variables

Tax rate
Net Value Per Share
EPS-Net Income
Debt ratio %: Liability/Total Assets
Borrowing dependency: Cost of Interest-bearing Debt
Operating profit per person: Operation Inc Per Employee
Working Capital to Total Assets
Current Liability to Assets
Working Capital/Equity
Current Liabilities/Equity
Retained Earnings to Total Assets
Total expense/Assets
CFO to Assets
Liability-Assets Flag
Net Income to Total Assets
Gross Profit to Sales
Net Income to Stockholder's Equity
Liability to Equity

Heat Map



- Heatmaps were used to identify multicollinearity among the variables in the model to ensure that it was not present
- There is some correlation, bit its not significant enough
- The common variables that are correlated to bankrupt for all the 3 datasets are used in 3 training datasets

train_og

train_boot

train_adas

Data Resampling

Re-sampling the Data

We seprated our test data in the same ratio as our original data.

We re-sampled our data to deal with the imbalance using the bootstrap with replacement and synthetic bootstrapping using Adaptive Synthetic Sampling (ADASYN) algorithm.

Data	No	Yes
Original Data	6599	220
Test Set	1500	50
Train set	5099	170
Train Boot	3000	500
Train ADSYN	5099	5152

Building Our Model

Logistic Regression

Classic model to predict the probability of a binary outcome

Decision Tree

Decision Trees tend to work well with imbalanced data

Model Errors-Logistic Regression

Model	Train Error	Class 1	Class 0	Test Error	Class 1	Class 0
Original Train	3.47%	61.17%	1.54%	3.87%	72%	1.6%
Bootstrapped Train	9.11%	45.8%	3%	4.70%	60%	2.86%
ADASYN Train	10.80%	10.34%	11.27%	11.2%	20%	11%

Model Errors–Decision Tree

Model	Train Error	Class 1	Class 0	Test Error	Class 1	Class 0
Original Train	3.05%	94.7%	0%	3.54%	100%	0.33%
Bootstrapped Train	8.05%	18.4%	6.33%	6.45%	38%	5.22%
ADASYN Train	10.57%	12.36%	8.76%	8.77%	34%	7.93%

Best Model

Logistic Regression with ADASYN Train

```
##
## Coefficients:
##               Estimate Std. Error z value             Pr(>|z|)
## (Intercept)   167.0427    19.5686   8.536 < 0.0000000000000002 ***
## X15             0.5194     0.2743   1.893              0.05829 .
## X18            -4.1505     2.9049  -1.429              0.15306
## X19           -61.3436     4.4711 -13.720 < 0.0000000000000002 ***
## X37            16.9427     2.2456   7.545   0.0000000000000453 ***
## X40           184.8304    12.5566  14.720 < 0.0000000000000002 ***
## X52            -2.2390     2.0327  -1.101              0.27069
## X54             9.7153     1.5556   6.245   0.0000000004231045 ***
## X60            16.5366     2.7843   5.939   0.0000000028618956 ***
## X65          -144.4739    15.0711  -9.586 < 0.0000000000000002 ***
## X66          -189.8534    38.7263  -4.902   0.0000009465525807 ***
## X68            -0.3500     2.8739  -0.122              0.90306
## X70            -8.0730     1.3768  -5.864   0.0000000045275152 ***
## X82            -1.8411     0.7157  -2.572              0.01010 *
## X85            10.6461     2.6403   4.032   0.0000552727583670 ***
## X86           -14.1992     2.9749  -4.773   0.0000018146094624 ***
## X89             6.7979     2.8602   2.377              0.01747 *
## X90           -35.2725    11.9687  -2.947              0.00321 **
## X91           -95.7023    34.5691  -2.768              0.00563 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significant Financial Variables

Tax rate
Net Value Per Share
Debt ratio %: Liability/Total Assets
Borrowing dependency: Cost of Interest-bearing Debt
Operating profit per person: Operation Inc Per Employee
Current Liability to Assets
Working Capital/Equity
Current Liabilities/Equity
Retained Earnings to Total Assets
CFO to Assets
Liability-Assets Flag
Net Income to Total Assets
Gross Profit to Sales
Net Income to Stockholder's Equity
Liability to Equity

Model	Train Error	Class 1	Class 0	Test Error	Class 1	Class 0
Logistic Regression with ADASYN Train	10.80%	10.34%	11.27%	11.2%	20%	11%

Recommendation

- Regularly monitor key financial metrics such as working capital to total asset/Equity, Net income to total asset, Total Expense /Asset, Debt ratio
- Using a predictive model like ours investors and shareholders can identify companies or start-ups that may be at risk of bankruptcy
- Maintain a health working capital to assests ratio to meet all its short term financing obligagtion.
- Maintain strong cash flow and keep debt under control. And also look into building reserve and diversifying their revenue streams

Thank You!

Q/A