Estimating and Visualizing Banks Failure using Random Forest

Vincenzo Dentamaro ,Daniel Cazzaniga, Thomas Weldon, Aswin Gigi, Sharath Kumar Ravi Kumar, Paul Livesey

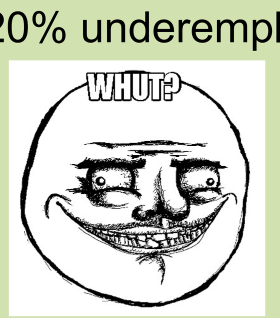
Problem, Solution and Literature Review

Problem:
Banks are the bedrock of our modern global economy, when they fail, the consequences can be both widespread and devastating even to ordinary people's life.

Looks Familiar?

How to avoid it? – Laymen's perspective

- If you're an investor in financial companies, it's simple: Stay away from names that look like the last Lehman Brothers: those with assets-to-equity ratios in excess of 15, that are funded largely with short-term borrowings, and whose arcane business models no human could possibly hope to understand.
- For everyone else, avoid banks who invests in companies that would be hurt by another financial meltdown, or who have jobs and would prefer not to see a return to 20% underemployment.

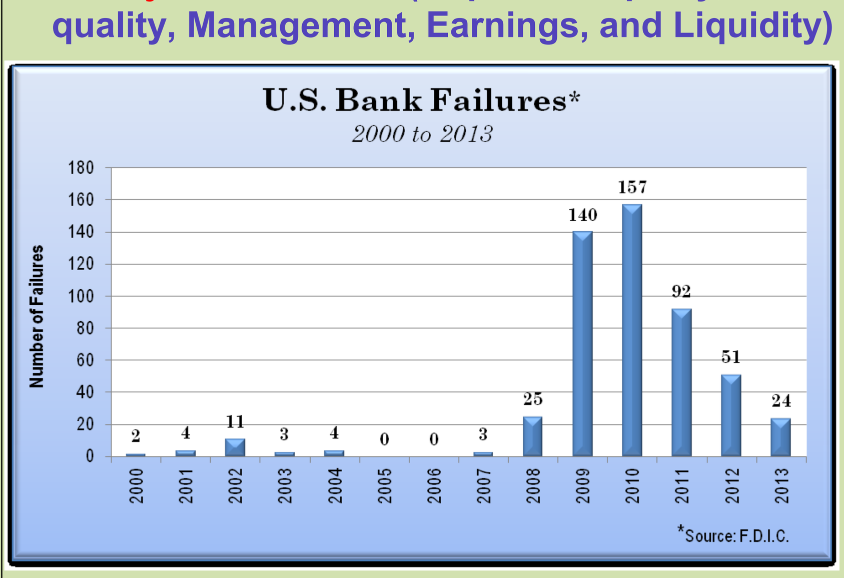


Statistical Approach

Since Laymen term fails to make common people analyze the risk, sometimes even to the trained eye, the better approach is statistics.

Statistics and data analysis:
Data is the modern fossil fuel. We already have huge collection of data regarding bank statistics. These data are used to visualize the cause of a bank failure or its prosperity.

***Keyword**: CAMEL (Capital adequacy, Asset quality, Management, Earnings, and Liquidity)*



Existing Methodologies to Predict Bank Failures

Most of the methodologies involve AI (Machine Learning / Neural networks)

- In 1968 Altman published a method of assigning a Z-score using a linear combination of five financial ratios with coefficients determined using Multiple Discriminant Analysis (MDA) with 92% accuracy
- A statistical model using probability theory able to show that the interest rate on a loan becomes a worse predictor of default as securitization increases. The paper didn't use any machine learning(ML) technique
- A ANN and self-organizing maps to display the probability of distress up to 3 years before bankruptcy. It uses 32 engineered features and used the Federal Deposit Insurance Corporation (FDIC)with an overall accuracy of 96.15%.

Our Solution:

This is the first attempt to develop a visual computer aided Financial Distress Predictor system making use of Random Forest ML Classifier on real data

Stages of development

Dataset
- FDIC bank records collection of USA from 2000 to 2018
- Visualization of dataset – clustering
- Labelling dataset

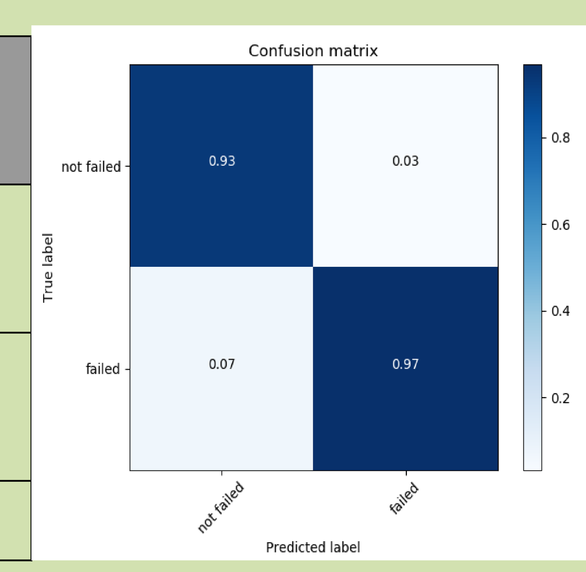
Predictor
- Splitting to Training set and test set
- Feature selection
- Training decision trees using Random forest classifier

Visualization
- Choropleth map of every county of USA
- Display Banks under risk when mouse hovers
- Search box to search for bank and pin code
- Decision tree visualization

Key Features
1. Asset Value
2. Change in balance
3. Change on interests
4. Percentage change in net loans and leases
5. Normalized Trade Value
6. Federal Fund sold
7. Normalized Premises and fixed assets
8. Normalized Intangible
9. Normalized Deposit
10. Normalized Deposit Interest
11. Normalized Total Domestic Deposit
12. Foreclosure ratio Income earned not collected on loans
13. bank size (log(asset))
14. Wholesale funding over asset
16. Other Liabilities over asset
17. Total equity-capital Percent change in non-current loans and leases
18. Total risk-weighted asset adjusted
20. Volatility of liabilities
21. Liquidity

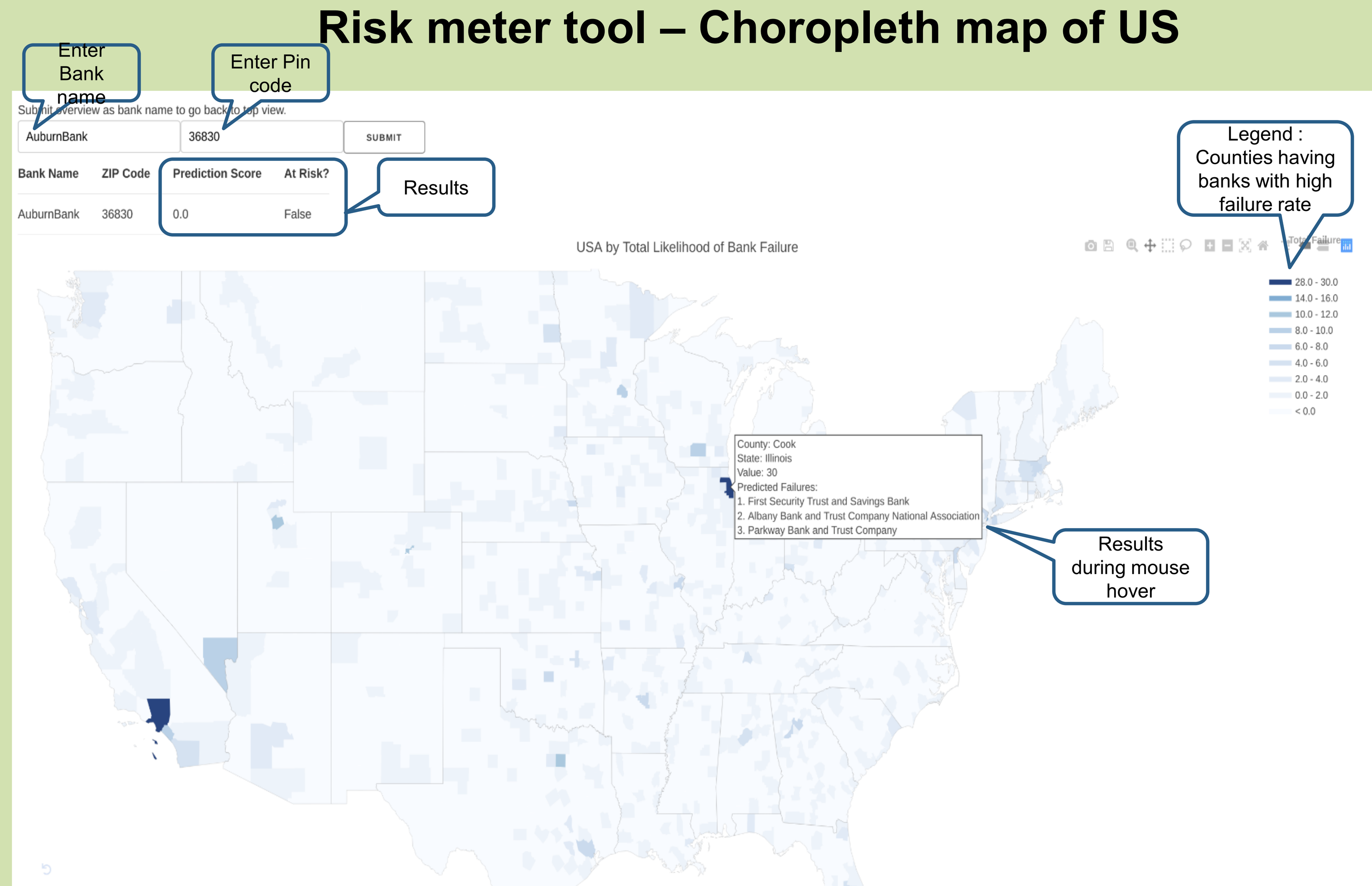
Outcome

Attribute	Results
Training Time	46 minutes
Accuracy	Training set – 98% Test set – 95%
F1 score	0.95



Highlights
- Dataset size is more than 500,000 rows of data collected for a period of 8 years on quarterly basis
- The dataset is very skewed with less than 3% of failed banks.
- Hence it was paramount to focus on True Negatives and to keep it as low as possible (we are able to achieve 7% of TN)
- We are able to achieve this by transforming a binary classification problem into a regression problem and later finding the correct threshold value to discern risky banks from non risky banks in the test set with 95% accuracy.

Risk meter tool – Choropleth map of US



Key Benefits:
1. Accurate FDP predictor superior to the literature reviewed (95% accuracy)
2. Interactive UI for searching non safe banks
3. The resulting decision-tree for rule based FDP modelling.
4. Graph showing similar banks clustered together allowing to explore hidden common patterns.
   In this way if a bank in the cluster has high probability of failure, then others in the same cluster need further deepening.
5. Also the choropleth map indicates that Illinois and California have more banks that are predicted to be at risk in 2019

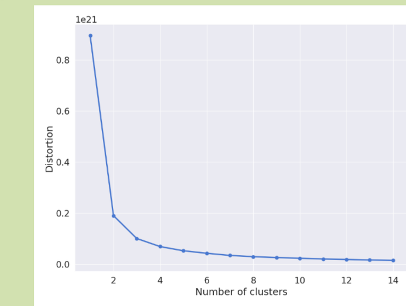
Further Improvements:

The risk meter tool needs to be simplified including also a description on the meaning of total failure.

The UI needs to be more responsive

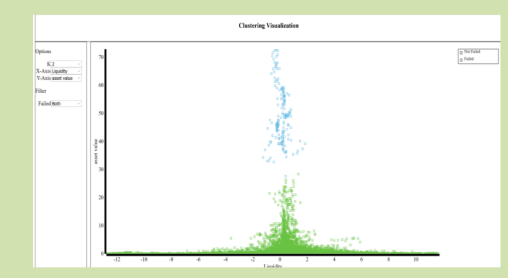
Dataset Visualization

Clustering:
- Data exploration by clustering is crucial to filter the features and to clearly classify the dataset
- For our work we have used K-means clustering with 3 clusters based on elbow graph and manual analysis



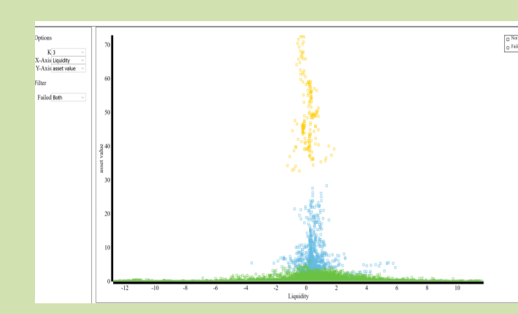
Two group Clustering:
The plot of two clusters, light blue and green over Liquidity (x axis) vs Asset Value (y axis).

Ideally we are able to classify the green Clusters as failed banks. But the results have more True Negatives



Three group Clustering:
The plot of three clusters, light blue, Orange and green over Liquidity (x axis) vs Asset Value (y axis).

In this clustering all Failure banks are grouped in green region and only 4 of the Failed banks are grouped in light blue region. All banks in orange region are safe



Decision tree Visualization

- The main problem in any ML Classifier is that it is a black box. Difficult to reason or understand the behavior of the algorithm
- But it is not the case for Decision trees as we are able to take out the exact decision made at each node.
- With the help of below Visualization graph it is easier to understand the decision points.