Predicting Company Bankruptcy

Grace Tang
Springboard DSC - Capstone 3

How do I know who to lend money to?

- A successful business loan can grow companies, create jobs, bolster the economy.
- A failed loan can result in millions of dollars in losses.

Picture Credit: debt.org

The Solution: Predict Ahead of Time

- Can we predict company bankruptcy, and use the results to inform financial decisions?
- Many applications:
  - Approving/Denying business loans
  - Investing more safely
  - Understanding what contributes to company health

The Data: 6819 Taiwanese Companies

Index	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	...
0	1	0.370594	0.424389	0.405750	0.601457	0.601457	0.998969	...
1	1	0.464291	0.538214	0.516730	0.610235	0.610235	0.998946	...
2	1	0.426071	0.499019	0.472295	0.601450	0.601364	0.998857	...
3	1	0.399844	0.451265	0.457733	0.583541	0.583541	0.998700	...
4	1	0.465022	0.538432	0.522298	0.598783	0.598783	0.998973	...

- 95 features
- 1 class label (1 for bankrupt, 0 for non-bankrupt)
- 6819 total companies
  - 220 (3.23%) bankrupt companies
  - 6599 (96.77%) non-bankrupt companies

Why Taiwan? Not so different from the U.S.A.

- The most technologically advanced computer microchip industry in the world
- Developed capitalist economy
- Most government firms privatized
- 7th-largest economy in Asia and 20th-largest in the world

Photo Credit: https://www.chinaimportal.com/blog/electronics-manufacturers-taiwan/

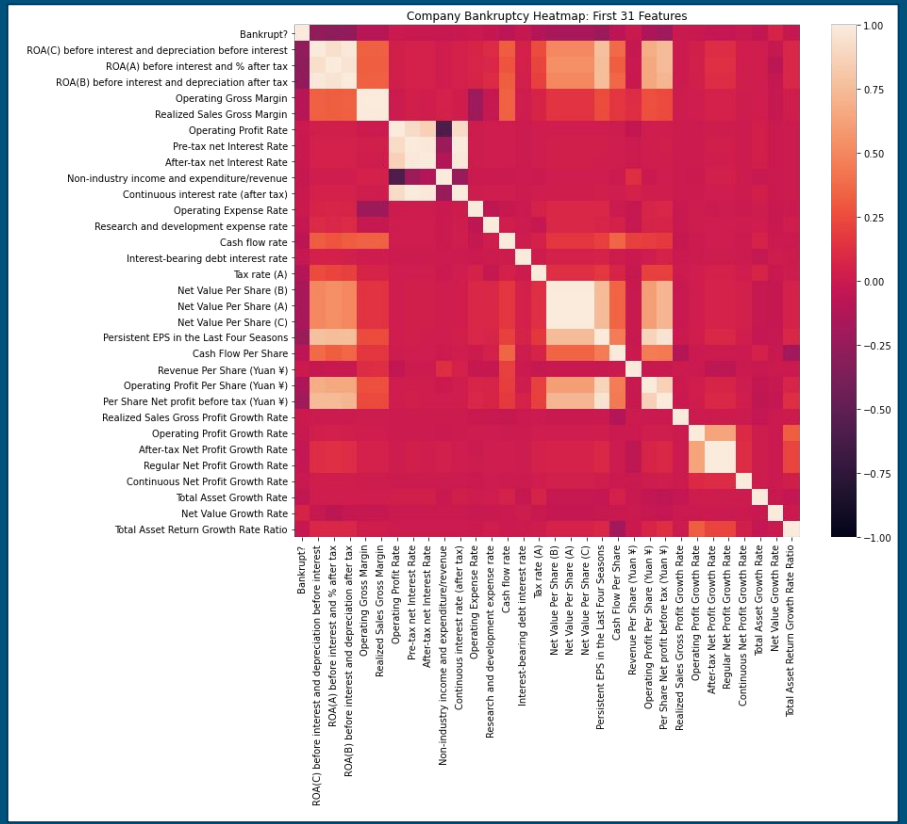
Which Features are Statistically Significant Between Class Labels? - Student's T-Test

Null Hypothesis: "There are no differences in feature values between 'bankrupt' & 'non-bankrupt' companies." (p-value < 0.01)

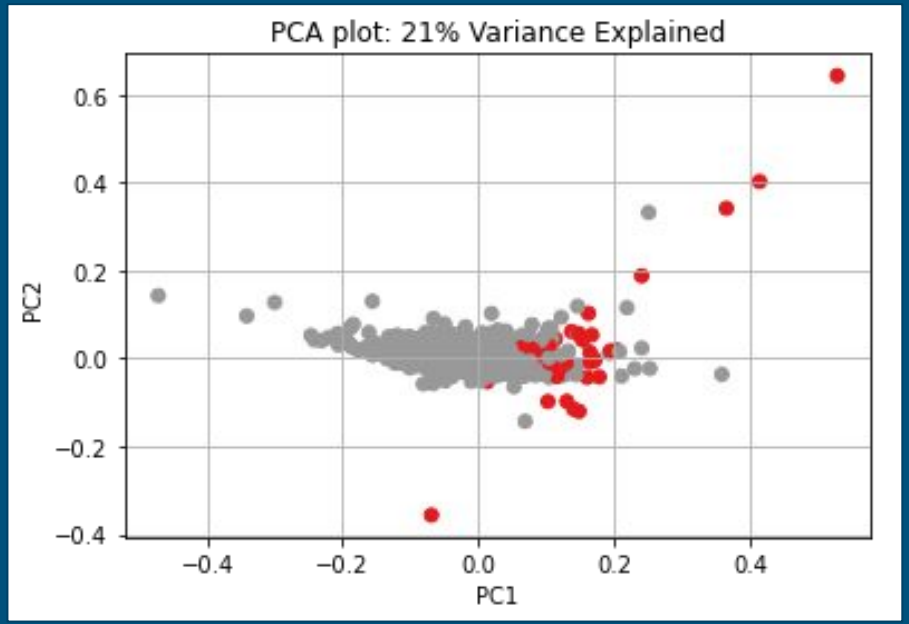
- 56 (60%) features significant
  - Return on Assets
  - Operating Gross Margin
  - Cash flow rate
  - Cash Reinvestment %
  - Debt ratio %
  - Net worth/Assets
  - …
- 38 (40%) features not significant
  - Operating Profit Rate
  - Pre-tax net Interest Rate
  - Non-industry income and expenditure/revenue
  - Operating Expense Rate
  - R&D expense rate
  - …

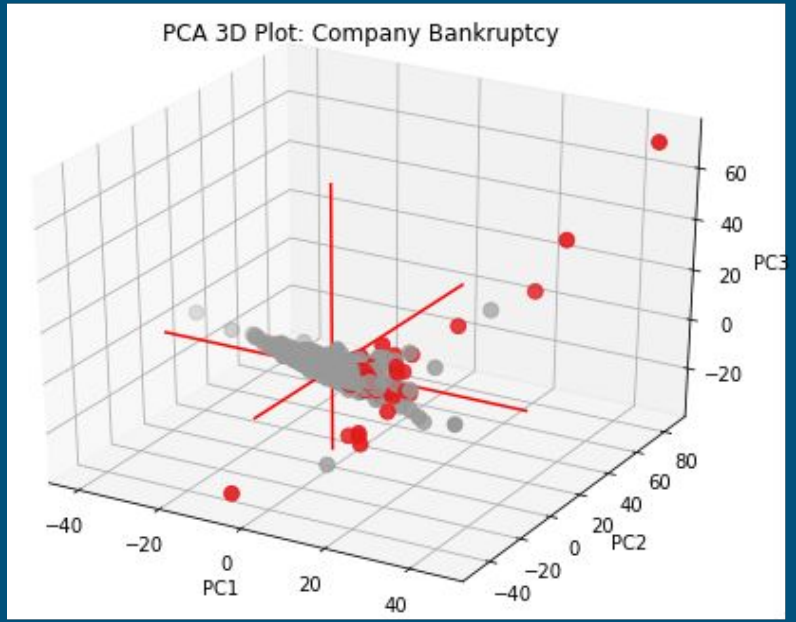
Features Non-Independent: Correlation Heatmap

- No strong correlations with bankruptcy.
- But some features have strong correlations with each other.
- Data would benefit from dimension reduction.



Dimension Reduction with PCA





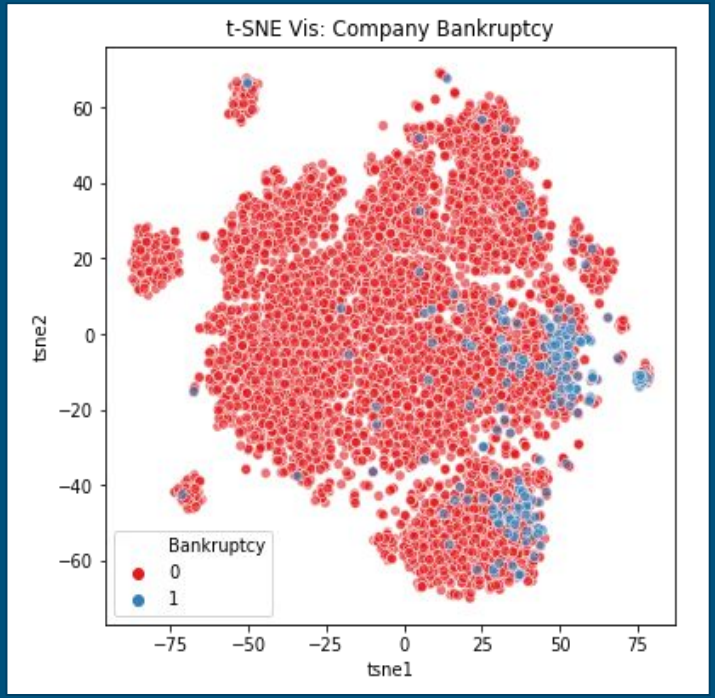
PCA: Features That “Reduce” Bankruptcy

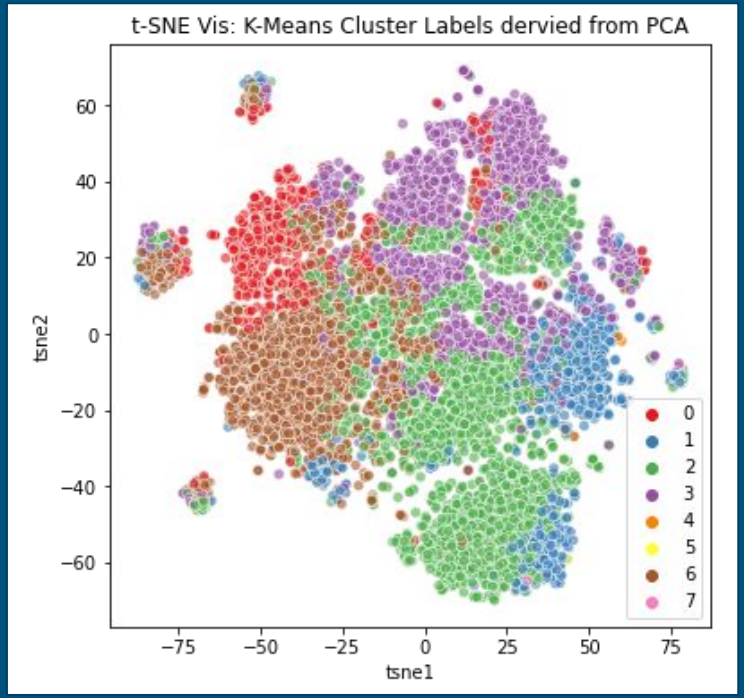
- Principal Component 1
  - Return on Assets
  - Persistent EPS in the Last Four Seasons
  - Operating Profit Per Share
  - Per Share Net profit before tax
  - Operating profit
  - Net profit before tax
  - Net Income Total Assets

PCA: Features That “Contribute” to Bankruptcy

- Principal Component 2
  - Debt ratio %
  - Borrowing dependency
  - Inventory and accounts receivable
  - Net Worth Turnover Rate (times)
  - Current Liability to Assets Ratio
  - Current Liability to Equity Ratio

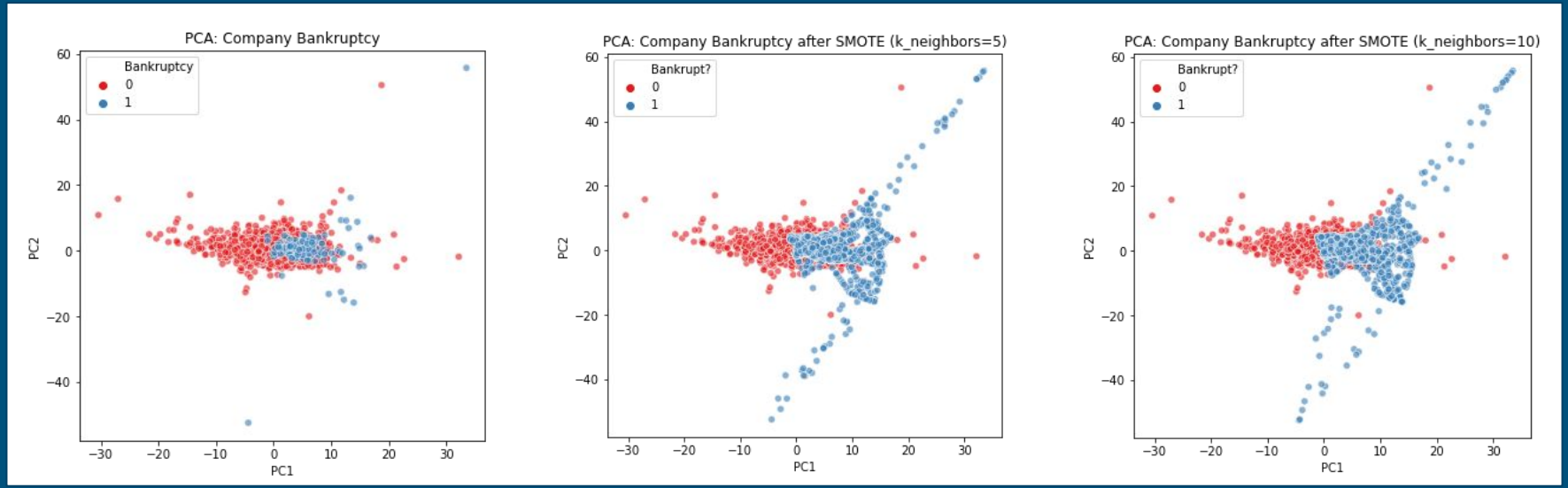
Visualizing with t-SNE and Cluster Labels





Upsampling the Minority (Bankrupt) Class

- Remember, only 3.23% is in the “Bankrupt” class.
- Upsample with SMOTE (Synthetic Minority Oversampling TEchnique)



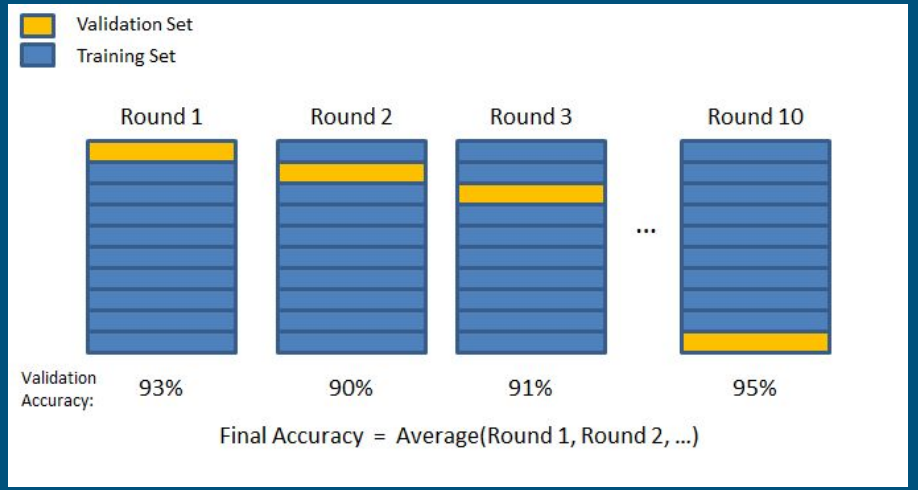
SMOTE Upsampling Improves Model Performance

Baseline Model: Logistic Regression		
Evaluation Metric	**Before SMOTE**	**After SMOTE**
Accuracy	0.9619	0.8773
F1-Score	0.0000	0.2807
Precision	0.0000	0.1731
Recall	0.0000	0.7424

Model Evaluation

- Binary Classification Models:
  - Logistic Regression
  - Random Forest
  - XGBoost

Hyperparameter Tuning: Random Search 10-Fold Cross Validation



Picture Credit: Chris McCormick, chrisjmccormick.wordpress.com/

Table of Results

Model	Data Preprocessing	Accuracy	F1-Score	Precision	Recall	ROC AUC	P-R AUC	Runtime (s)
Logistic Regression	No	0.9619	0.0000	0.0000	0.0000	0.5887	0.0377	1.0776
Logistic Regression	Yes	0.8739	0.2989	0.1821	0.8333	0.9252	0.3020	2.0035
Random Forest	No	0.9707	0.2308	0.7500	0.1364	0.9497	0.4728	1.0542
Random Forest	Yes	0.9355	0.2979	0.2295	0.4242	0.9203	0.2473	6.6547
XGBoost	No	0.9697	0.3404	0.5714	0.2424	0.9472	0.4602	0.3819
XGBoost	Yes	0.9355	0.3465	0.2574	0.5303	0.9086	0.2524	6.2850

Takeaways:

- F1-Score is higher for the models fitted to the processed data.
  - scaled, dimension reduction, upsampling with SMOTE
- Logistic Regression Model with data preprocessing has highest recall (minimizes false negatives).
- XGBoost with data preprocessing has highest F1-Score.
  - XGBoost also more robust to outliers than LR.

Future Directions

- Understand the Model:
  - Investigate the underlying features that contribute to bankruptcy. (Build on results from PCA and t-test.)
- Improve the Model:
  - Further investigation into data preprocessing methods to improve separation between classes.
- Apply the Model:
  - Implement model pipeline on USA business data and evaluate.

Thank you

Chris Esposo for being an awesome Springboard mentor: for teaching me about SMOTE, helping me narrow down ideas, and giving me guidance throughout this project.

Deron Liang and Chih-Fong Tsai (National Central University, Taiwan) for donating this valuable dataset to the UCI Machine Learning Repository.

Federico Soriano Palacios for uploading the dataset to kaggle, where I was able to access it.

Citations

Dataset donated by:
Deron Liang and Chih-Fong Tsai, deronliang '@' gmail.com; cftsai '@' mgt.ncu.edu.tw, National Central University, Taiwan

Relevant Paper:
Liang, D., Lu, C.-C., Tsai, C.-F., and Shih, G.-A. (2016) Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study. European Journal of Operational Research, vol. 252, no. 2, pp. 561-572.

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

Questions?