Predicting Corporate Bankruptcy

Derek Dewald
Guanghua Li
Varun Naidu

Berkeley
UNIVERSITY OF CALIFORNIA

Intro

- Objective: Develop a model to predict bankruptcy risk in Polish companies
- Importance: Financial stability, risk management, and resource allocation
- Business Applications: Risk assessment, credit extension, supplier selection(Banking, insurance, supply chain management)

Dataset

Facts

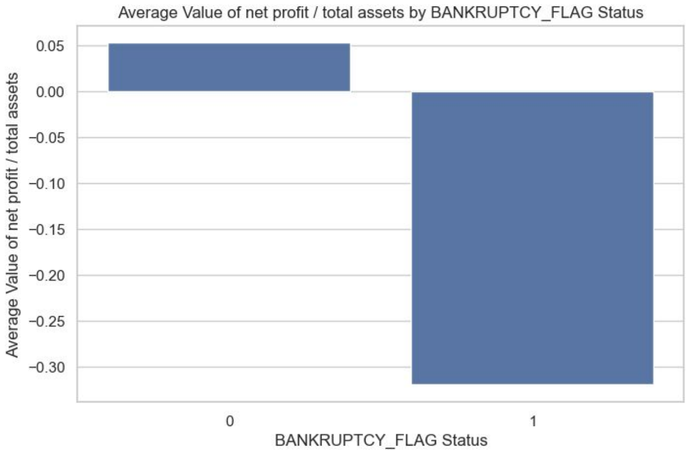
- Polish Corporate Entities
- 60+ Financial Ratio
- 5 Years of Observations 2008 – 2012
- ~8000 Records per Annum
- Bankruptcy Evidenced ~5%

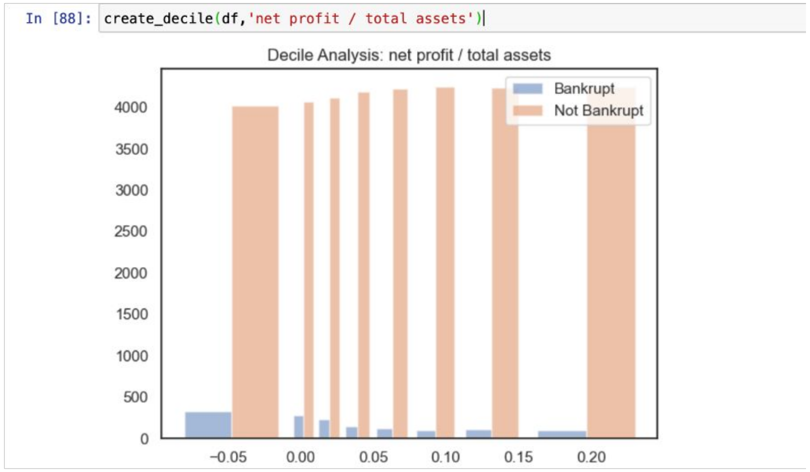
Observations

- Types of Companies Unknown
- Potential Multicollinearity
- What does "Bankruptcy" imply and how does it impact our design

EDA - Overview

- Histogram of all 60 Variables
  - Standard Deviation High
  - Outliers prevent simple visualization
- Decile Analysis
  - Insight to Correlation
  - Enabled Comparison
- Explore Financial Ratios
  - Liquidity
  - Capitalization
  - Turnover
  - Profitability



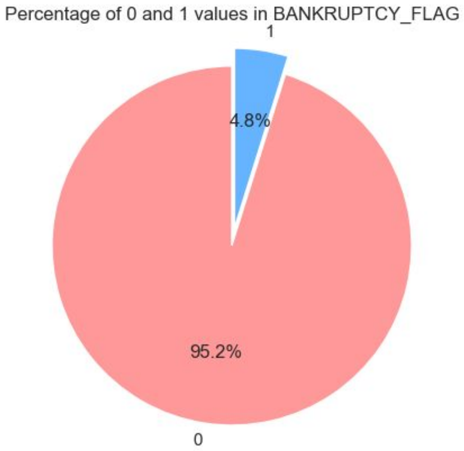


EDA - Sample

- Explore Data Quality
  - Initial Approach Review Columns
    - Remove
    - Clean
    - Impute

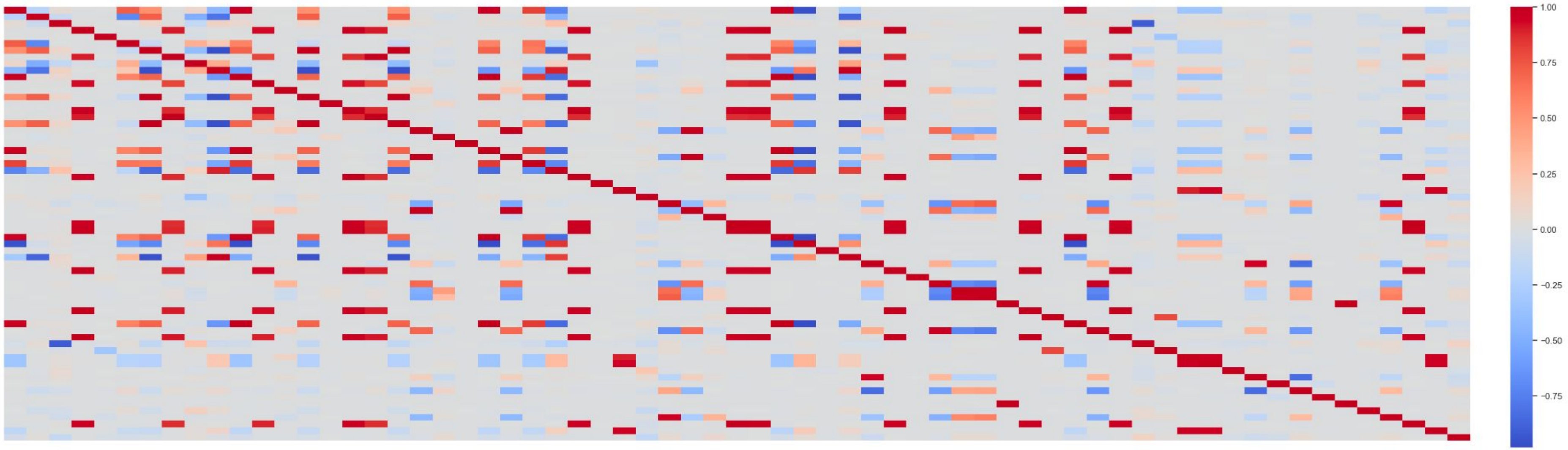
```
Statistics for 'Net Profit / Total Assets' column:
Number of null values: 8
Number of zeros: 240
count    43397.000000
mean         0.035160
std          2.994109
min       -463.890000
25%          0.003429
50%          0.049660
75%          0.129580
max         94.280000
Name: net profit / total assets, dtype: float64
```

- Imbalanced Dataset
  - Oversample or downsample



EDA - Correlation

- Some variables are strongly correlated
- Correlation amongst variables appears to be a concern
- Strongly correlated variables could be redundant for the model



Data Cleaning and Other Challenges

Data Challenges

- Systematic Removal: Incomplete financial records are identified and removed based on null values in critical financial ratios.
- Adjustment of Ratios: Negative financial ratios are adjusted to zero to ensure they reflect realistic financial conditions and interpretations.
- Data Integrity: The cleaning process enhances the dataset's reliability, enabling more accurate financial analysis and assessments.

Other Challenges

- Company Industry not available. impact of relevance of financial ratios
- Company size largely obfuscated by ratios
- 2008 Financial Crisis – subsequent bail outs, structural change?
- Quantitative Information – Leadership, experience, organizational culture
- Other causes of Bankruptcy

Features

- Based on correlation with bankruptcy we have identified 7 variables, which we will prioritize as tier 1.

- These variables meet both an intuitive reference with what is believed a pragmatic model and appears to mathematically be relevant

	Financial Ratio	Ratio Classification
0	net profit / total assets	Profitability Ratio
1	total liabilities / total assets	Capitalization Ratio
2	working capital / total assets	Liqudity Ratio
5	retained earnings / total assets	Capitalization Ratio
28	logarithm of total assets	Capitalization Ratio
50	short-term liabilities / total assets	Capitalization Ratio
54	working capital	Liqudity Ratio

Baseline Analysis

- Identified Tier 1 – Variables
  - Strong correlation
  - Consistent with intuition

Baseline Model	1) Bankruptcies Predicted	2) True Positives	3) True Negatives	4) False Positives	5) False Negatives	6) Precision	7) Recall	8) Accuracy
NOT_PROFITABLE	9531	958	32733	8573	1132	2.84%	10.05%	77.64%
NO_LIQUIDITY	9612	903	32597	8709	1187	2.70%	9.39%	77.20%
LIABILITIES_GT_ASSETS	2277	307	39336	1970	1783	0.77%	13.48%	91.35%
ST_OBLIGATIONS_GT_TOTAL_ASSETS	1409	228	40125	1181	1862	0.57%	16.18%	92.99%
NO_EQUITY	27656	1575	15225	26081	515	9.38%	5.69%	38.71%
TWO_BINARY_FLAGS	7533	539	34312	6994	1551	1.55%	7.16%	80.31%
THREE_BINARY_FLAGS	2688	350	38968	2338	1740	0.89%	13.02%	90.60%
FOUR_BINARY_FLAGS	730	71	40647	659	2019	0.17%	9.73%	93.83%
FIVE_BINARY_FLAGS	1013	179	40472	834	1911	0.44%	17.67%	93.67%
ALWAYS_BANKRUPT	43396	2090	0	41306	0	100.00%	4.82%	4.82%
NEVER_BANKRUPT	0	0	41306	0	2090	0.00%	0.00%	95.18%

Berkeley
UNIVERSITY OF CALIFORNIA

Baseline Analysis Cont.

- Imbalanced Dataset
  - 95% Accuracy. 0% Recall, 0% Precision

- Where does Model Value Lie?
  - Clear purpose statement, evaluation criteria

- Maximize Expected Return on Capital
  - Develop a simplistic representation of expected value of model, based on ability to maximize interest revenue, while lowering loan loss provisions.

Approach 1: Financial Theory Led

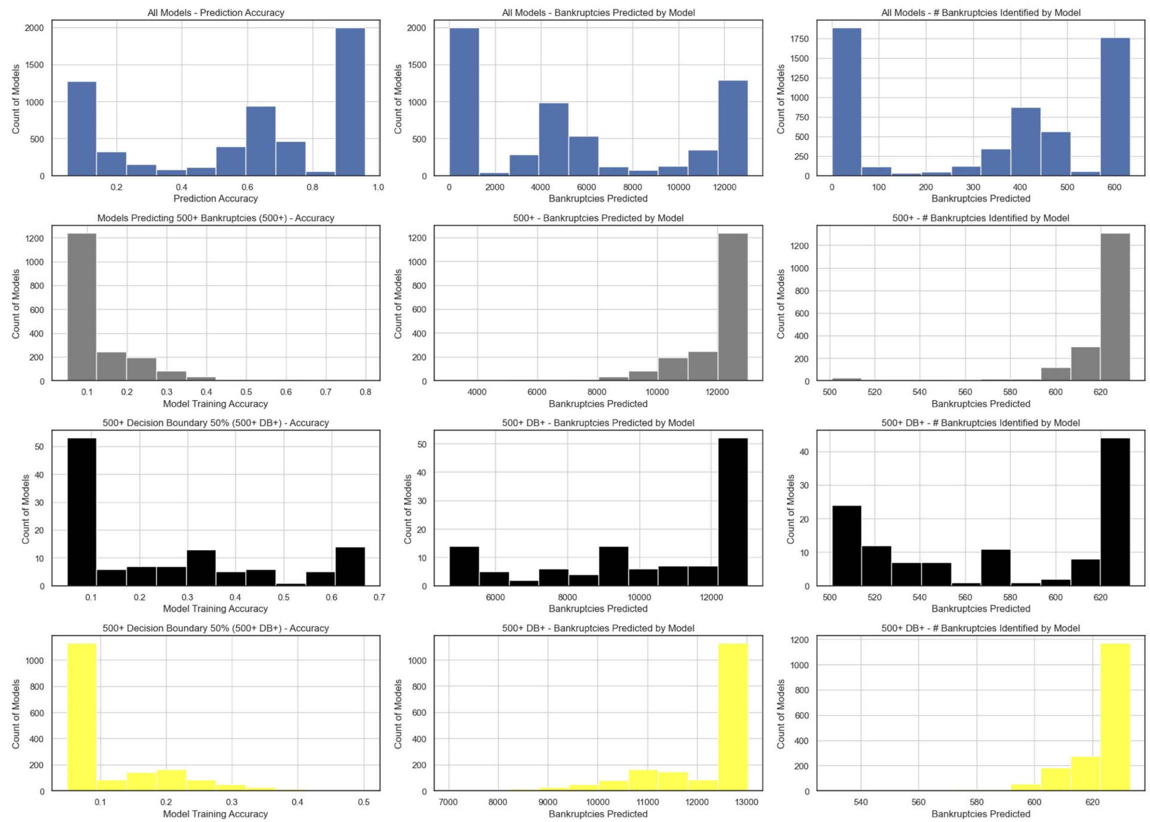
Data Pipeline Created

- Select Dataset
  - All Ratios, Tier 1 Ratios, Tier 1 Binary Ratios
- Select Data Standardization
  - None, Min Max Scaler, Standard Scalar
- Select Size of Data
  - All Data, 1000 Balanced Entries, 1500 Balanced Entries
- Select Unique Model Parameters(Ie. Neural Network)
  - Activation functions: relu, tanh, sigmoid
  - Optimizers: Adam, SGD
  - Batch Size: 100,1000
  - Learning Rate: .01, .05
  - Epochs: 10,20
  - Layer Sizes; [[8],[8,16],[8,16,32],[8,16,32,64]],
  - Decision Boundaries:[20,30,40,50],

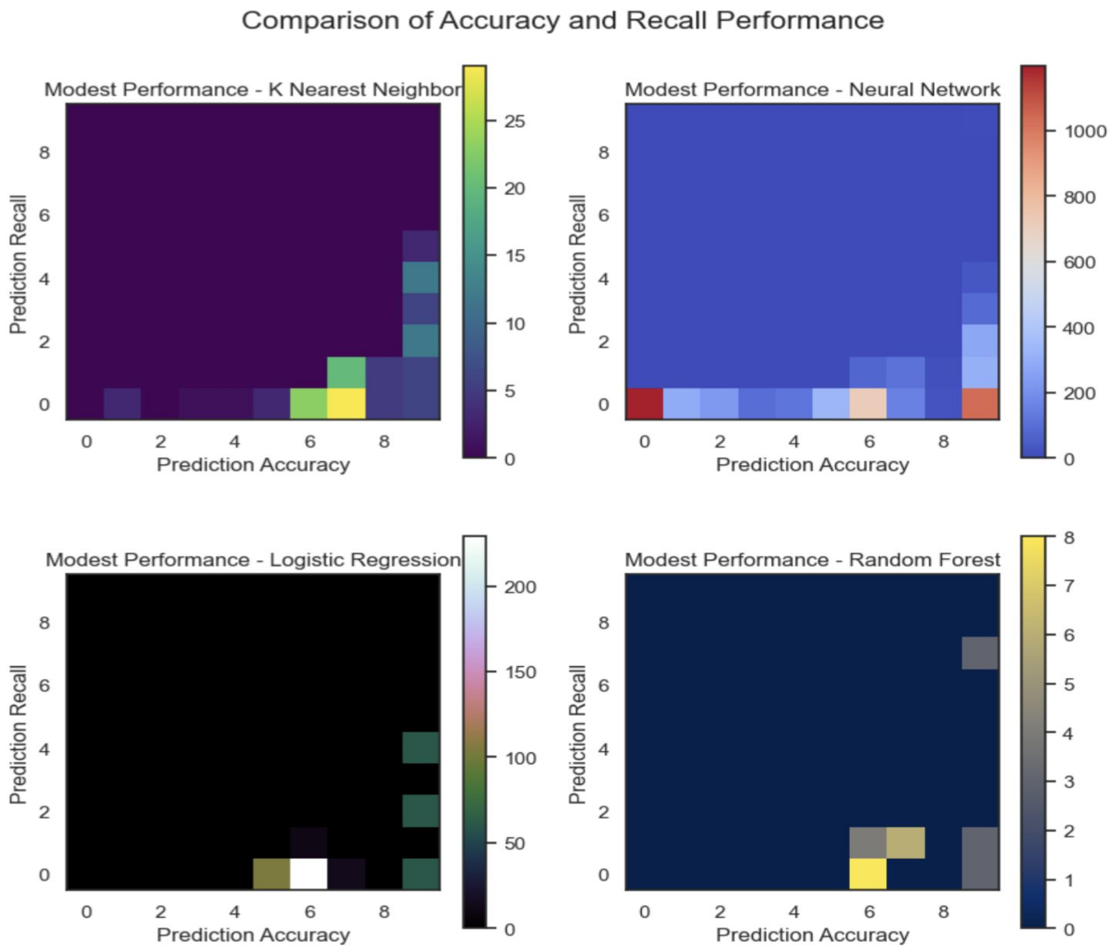
Data Experiments and Models

- ~5800 Models Generated

- Neural Network
- Logistic Regression
- K Nearest Neighbors
- Decision Tree
- Proceed with caution due to Interpretability



Data Experiments and Models



Berkeley
UNIVERSITY OF CALIFORNIA

Approach 2 - Domain Agnostic

The dataset is divided into 5 parts based on the forecasting period (1 to 5 years)

- Each forecasting period contains financial ratios from that year and a class label that indicates bankruptcy status after (6 – forecasting period) years.
  - Ex: Class labels of Forecasting Period 3 indicate bankruptcy after 3 years
- Considering the temporal angle of this dataset –
  - 5 models for 5 forecasting periods
  - Individual processing of each dataset

Data Issues

- Each dataset had some features with a high percentage (> 20%) of missing values.
- Some features had high outliers.
- For each forecasting period, the dataset was highly imbalanced (93% to 96% of the observations belonged to the Did not go Bankrupt class) .
- The scale of features were highly varied.

Data Cleaning and Preprocessing

Cleaning

- Columns that had more than 20% missing values were dropped.
- High outliers were removed conservatively to avoid losing out on observations that indicate bankruptcy (Methods: z-score, isolation forest).

Preprocessing

- Each dataset was split into training (70%) and testing (30%) sets (observations of each class label were sampled individually and then combined).
- The missing values were imputed separately for training and test sets (knn, mean).
- Each training set was balanced (upsampling, downsampling, and smote).

Models Evaluated

- Baseline (Always predict that a firm did not go bankrupt)
- K-Neighbors
- Decision Tree
- Random Forest
- Gradient Boosting
- Histogram Gradient Boosting
- Neural Network

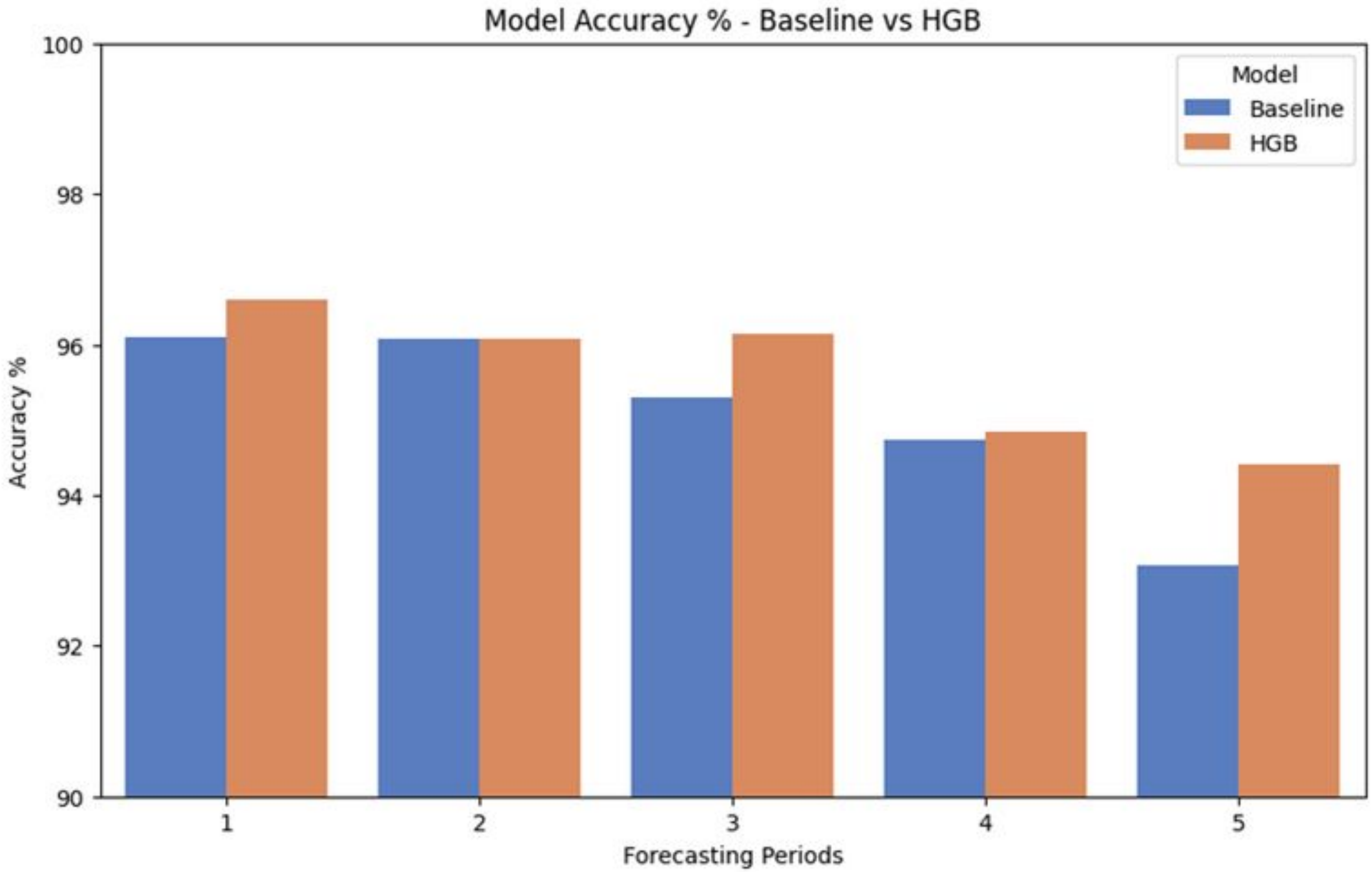
Best performing model - Histogram Gradient Boosting Classifier

- Hyper parameters tuned – Outlier removal by Isolation Forest, KNN Impute, Standard Scaling, Decision threshold, HGB max iterations, and HGB max depth.

Forecasting Period	Accuracy %	Class 0 Precision	Class 0 Recall	Class 1 Precision	Class 1 Recall	Threshold	HGB Max Iterations	HGB Max Depth
1	96.59	0.98	0.98	0.57	0.51	0.3	1000	3
2	96.07	0.98	0.98	0.50	0.44	0.2	1000	4
3	96.14	0.98	0.98	0.61	0.51	0.5	1000	3
4	94.84	0.97	0.98	0.51	0.36	0.2	1500	5
5	94.42	0.96	0.98	0.64	0.45	0.5	1000	3

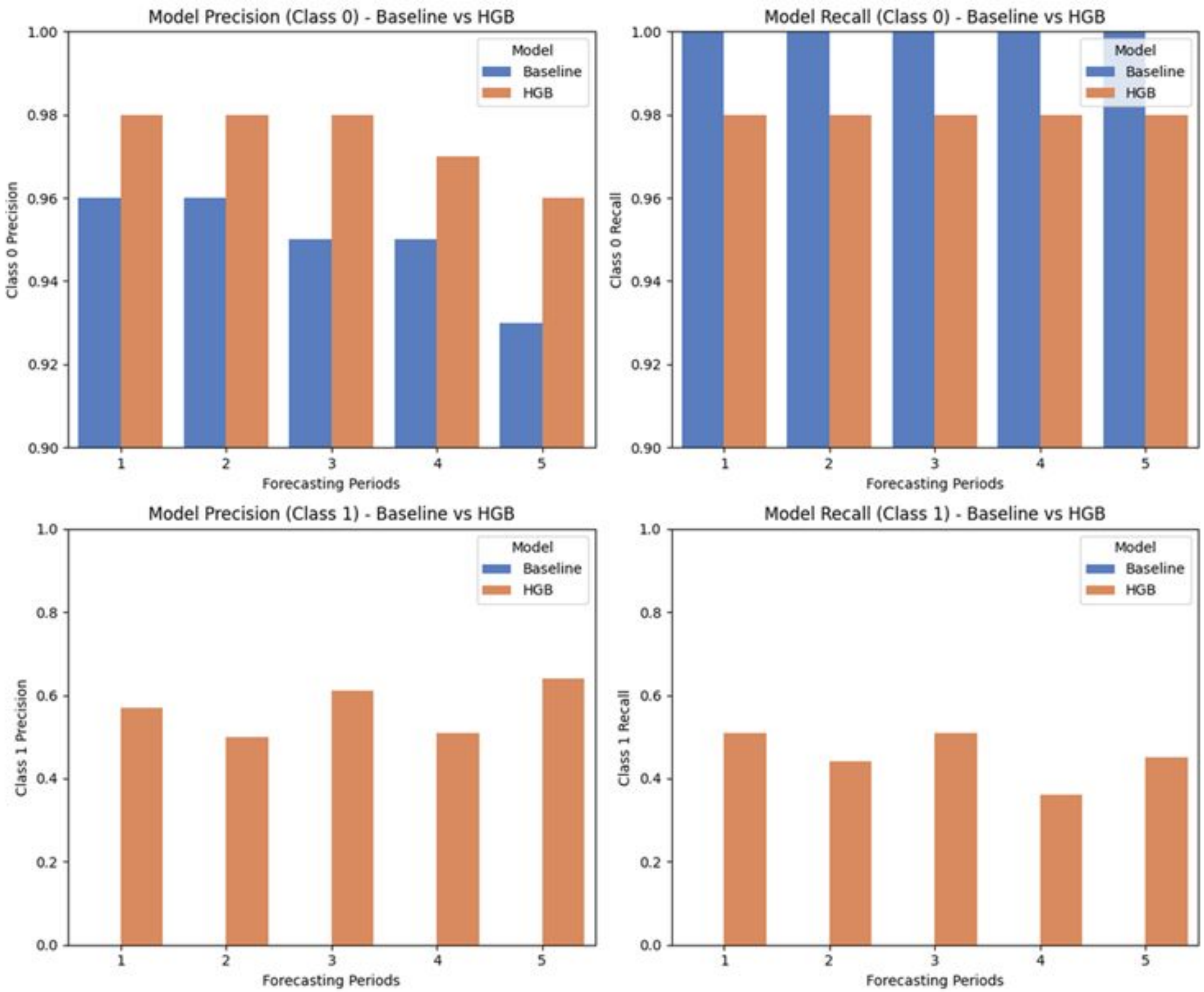
Berkeley
UNIVERSITY OF CALIFORNIA

Accuracy – Baseline vs Histogram Gradient Boosting Classifier



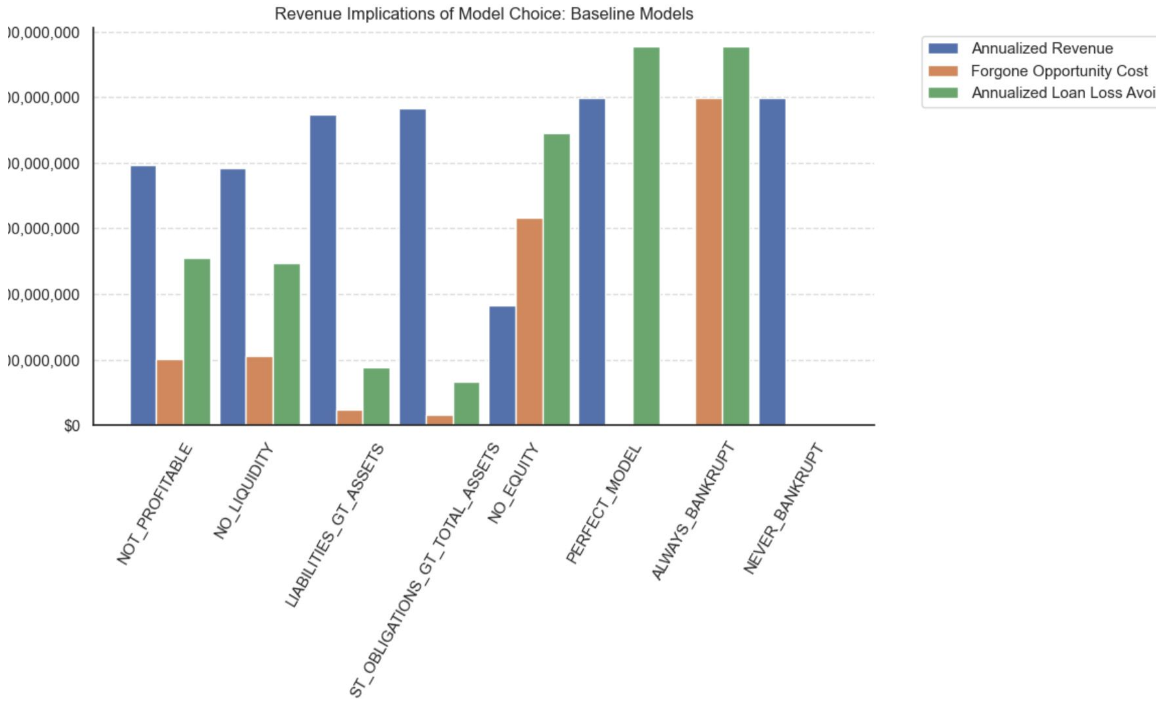
Berkeley
UNIVERSITY OF CALIFORNIA

Precision and Recall – Baseline vs Histogram Gradient Boosting Classifier



Berkeley
UNIVERSITY OF CALIFORNIA

Evaluation Criteria



- Predictability vs Interpretability
- Definition of "Bankrupt"
- Short Term vs Long Term Default
  - Model time matters
- Impact of Macro Economic Factors
  - 2008 financial crisis
- Trade off between revenue, opportunity cost and loan losses
- Different models satisfy different users, as such the work we did servers as a baseline to create a flexible and dynamic product which would be very attractive for financial institutions, venture capitalists, hedge funds, private equity and potentially others

Conclusion

- Success?
  - Model beat baseline from combined perspective of accuracy, precision and recall perspectives
  - Increase potential revenue, and reduce expected annual losses relative to baseline
- Best Model?
  - In the eye of the beholder
  - Related to the available data
- Next Steps?
  - Extend to other countries, test ability to generalize
  - Extensive data gathering to improve performance
  - Apply to very explicit and specific applications
- Data Limitations
  - As discussed, very real and challenging
- Approach Limitations
  - Financial engineering, financial fraud, availability of information

Appendix: Confirmation of Contribution

Topic	Derek	Guanghua	Varun
Code and Presentation	Primary contributor to Approach 1, including analysis, EDA, and conclusion supporting.		Worked on Approach 2, Data exploration, preprocessing, models and results.
Dataset and EDA	Contributed on Approach 1.		Approach 2
Approach and Models	Contributed on Approach 1.		Approach 2
Tuning and Improvements	As per approach 1, ran substantial data pipeline, which was primarily automated and enabled review of ~ 5800 combinations.		Approach 2 Turing of 6 models, improvement of best performing model
Conclusion and Checklist	Joint contribution with group.		Joint contribution with group.
Other Contributions	Github, Branch Derek Workbook DATASCI207_PROJECT_APPROACH1, which which walks through approach, EDA and more expansive details on logic. Colab – Need to Upload Files to your drive to replicate.		Github branch – varun Workbook – W207__Approach__2.ipynb (All steps associated to approach 2)

Berkeley
UNIVERSITY OF CALIFORNIA

NeurLPS Checklist

1. For all authors...
    - (a) Do the **main claims** made in the abstract and introduction accurately reflect the paper's contributions and scope? - Yes
    - (b) Have you read the **ethics review guidelines** and ensured that your paper conforms to them? - Yes
    - (c) Did you discuss any potential **negative societal impacts** of your work?. - Yes
    - (d) Did you describe the **limitations** of your work? - Yes
2. If you are including theoretical results...
    - (a) Did you state the full set of **assumptions** of all theoretical results? - NA
    - (b) Did you include complete **proofs** of all theoretical results? - NA
3. If you ran experiments...
    - (a) Did you include the code, data, and instructions needed to **reproduce** the main experimental results (either in the supplemental material or as a URL)? - Yes
    - (b) Did you specify all the **training details** (e.g., data splits, hyperparameters, how they were chosen)? - Yes
    - (c) Did you report **error bars** (e.g., with respect to the random seed after running experiments multiple times)? - NA
    - (d) Did you include the amount of **compute** and the type of **resources** used (e.g., type of GPUs, internal cluster, or cloud provider)? - NA
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
    - (a) If your work uses existing assets, did you **cite** the creators? - NA
    - (b) Did you mention the **license** of the assets? - NA
    - (c) Did you include any **new assets** either in the supplemental material or as a URL? - NA
    - (d) Did you discuss whether and how **consent** was obtained from people whose data you're using/curating? - NA
    - (e) Did you discuss whether the data you are using/curating contains **personally identifiable information** or **offensive content**? - NA

Berkeley
UNIVERSITY OF CALIFORNIA