Bankruptcy Prediction

Anthony Tagliente

Can we predict if a company will go bankrupt in the next year?

Sources:

- SEC EDGAR Database
- UCLA-LoPucki Bankruptcy Research Database
- Over 50,000 Annual and Quarterly filings

Collecting the data

311 bankrupt companies matched with CIK ids, resulting in over **4000 balance sheet** fillings from 2013 to 2020.

Inconsistent naming conventions and notes in headers lead to **thousands of individual columns** in our bankruptcy dataset.

Use the most common column names as base line.

By searching for columns with closely matching names across the set, we can condense the columns to about **57 useful features.**

Collecting the data

We use the same approach on NYSE companies not in the bankruptcy data set and perform the same data collection.

Giving us around **47,000 filings** for non-bankrupt companies.

The target variable is set to SEC filings dated within **365 days** of the company filing for bankruptcy.

Problems to account for

- KNN
- Logistic Regression
- Naive Bayes
- SVM
- RandomForest
- XGBoost

- Missing Values
- Large Class Imbalance
- Ambiguous Columns

Problems to account for

- ~~KNN~~
- ~~Logistic Regression~~
- ~~Naive Bayes~~
- ~~SVM~~
- RandomForest
- XGBoost

- ✔ Missing Values
- Large Class Imbalance
- Ambiguous Columns

Problems to account for

- ~~KNN~~
- ~~Logistic Regression~~
- ~~Naive Bayes~~
- ~~SVM~~
- RandomForest
- XGBoost

- ✔ Missing Values
- ✔ Large Class Imbalance
- Ambiguous Columns

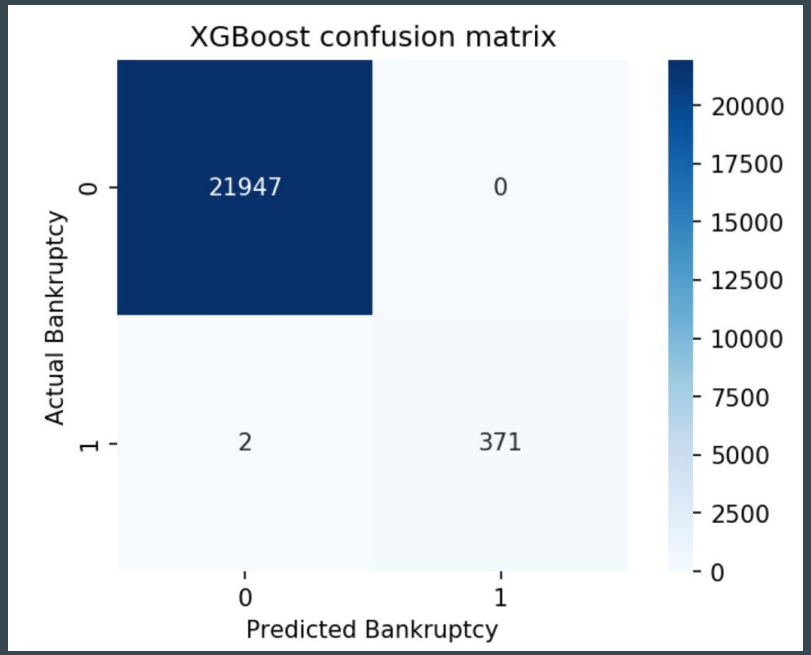
Problems to account for

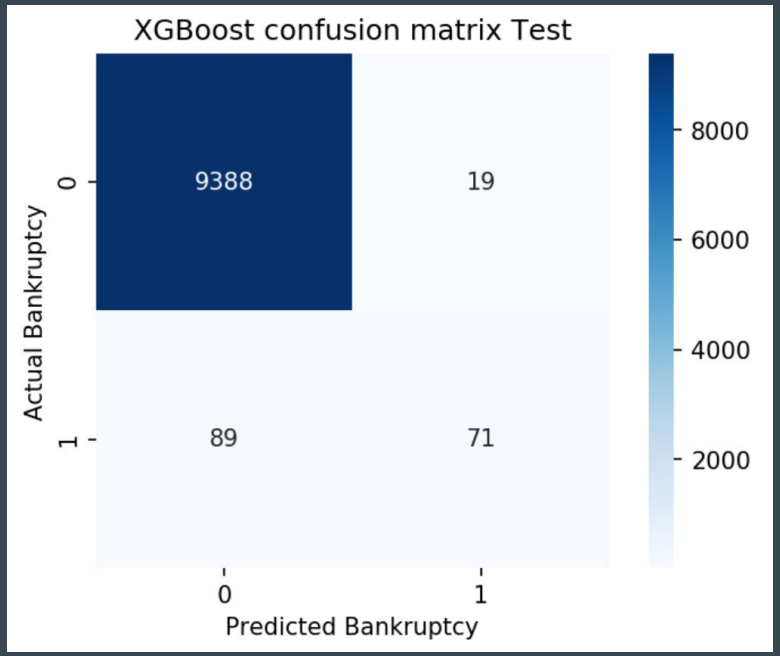
- ~~KNN~~
- ~~Logistic Regression~~
- ~~Naive Bayes~~
- ~~SVM~~
- RandomForest
- XGBoost

- ✔ Missing Values
- ✔ Large Class Imbalance
- ✔ Ambiguous Columns

Baseline Model

Recall: .44

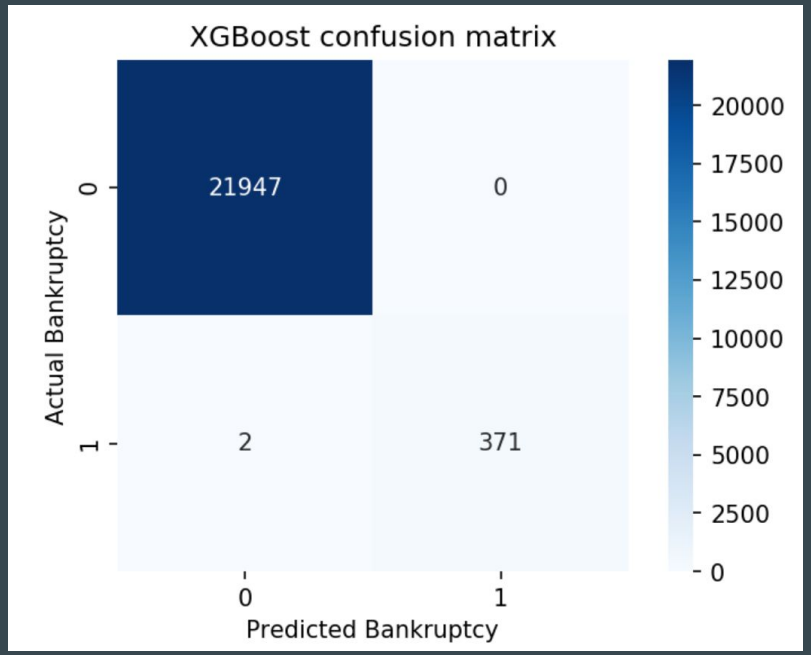


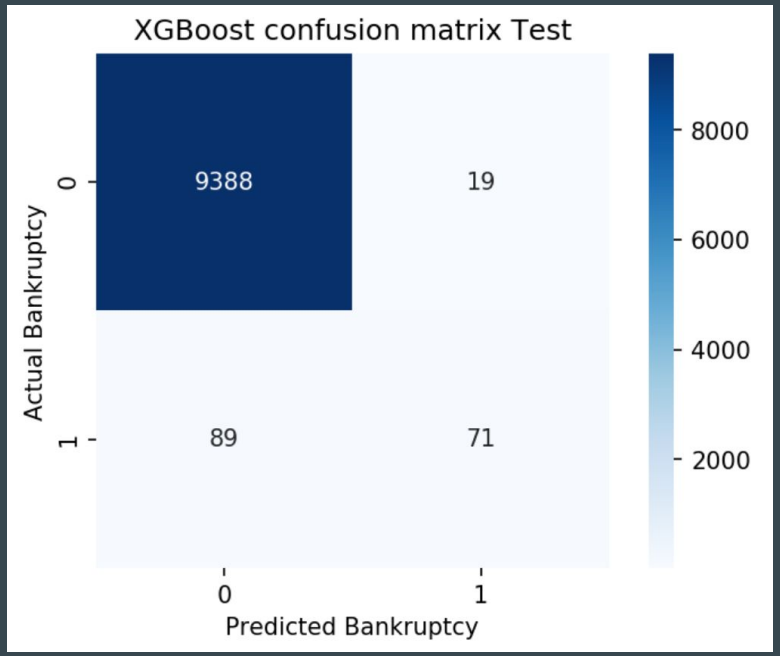


Baseline Model

We are onto something

Recall: .44

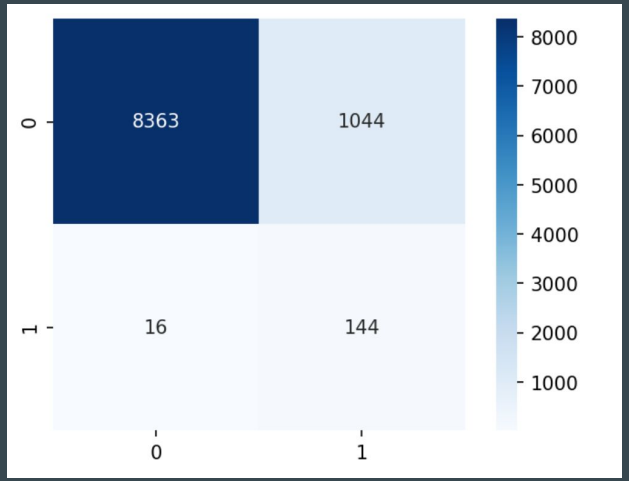




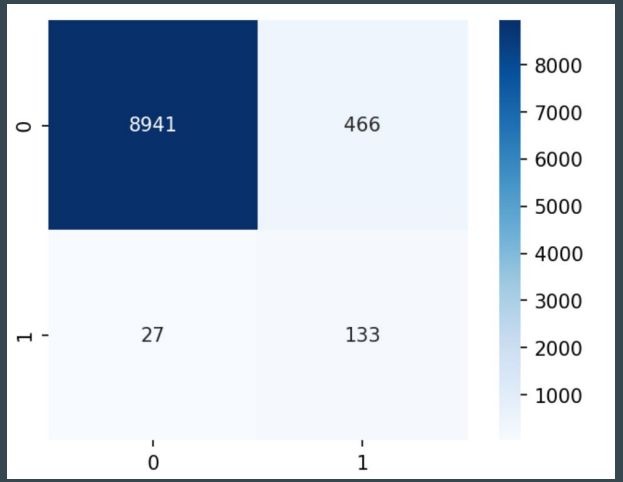
Synthetic Minority Oversampling Technique

- We have to impute for NaN values.
- Oversample bankruptcy class with synthetic data.
- Undersample majority class with random undersampling.

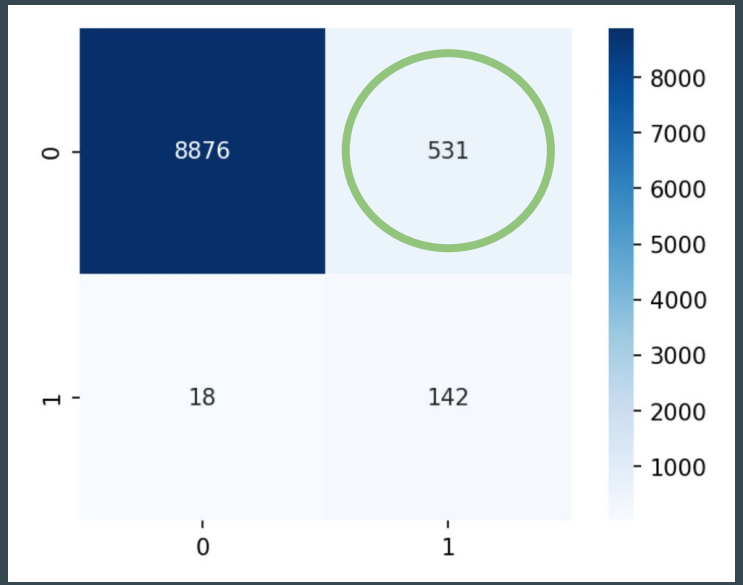
50/50 split Recall: .9



90/10 split w/ undersample: Recall: .83

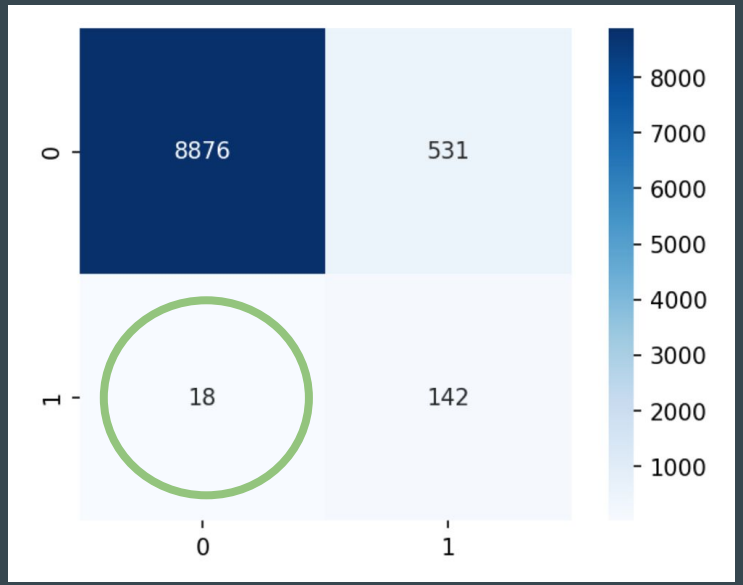


Hyperparameter tuning and adjusting sampling



- Of the 531 False positives 407 did file for bankruptcy eventually.

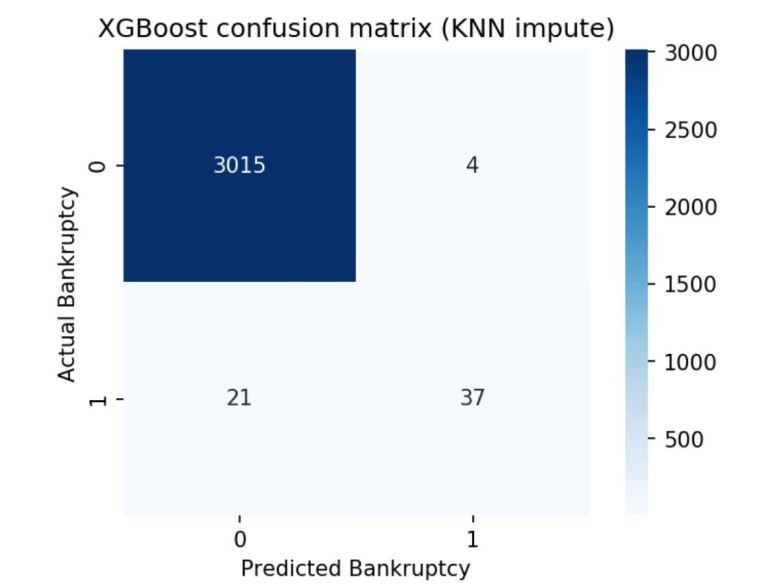
Hyperparameter tuning and adjusting sampling



- Of the 531 False positives 407 did file for bankruptcy eventually.
- **Of the 20 (2 in test) false negatives about 8 had been previously flagged.**

	CIK	Name	Filings	Filing Date	prdicted	banruptcy_6_mo	d_to_bankruptcy
21282	1168213	Aeropostale, Inc.	10-K	2014-04-04	0	0	761.0
21284	1168213	Aeropostale, Inc.	10-Q	2014-06-09	0	0	695.0
21285	1168213	Aeropostale, Inc.	10-Q	2014-09-08	0	0	604.0
21286	1168213	Aeropostale, Inc.	10-Q	2014-12-08	0	0	513.0
21283	1168213	Aeropostale, Inc.	10-K	2015-03-30	0	0	401.0
21287	1168213	Aeropostale, Inc.	10-Q	2015-06-11	1	1	328.0
21288	1168213	Aeropostale, Inc.	10-Q	2015-09-08	0	1	239.0
21289	1168213	Aeropostale, Inc.	10-Q	2015-12-07	1	1	149.0

Out of sample (2020) test



- Similarly, the model missed some reports after having correctly identified financial troubles in other periods.
- There is potential that some 2020 information **bled into the model.**
- The model seems to do **worse with reporting periods very close to bankruptcy** filing dates.

Making future improvements

- The model suffers from a general **lack of data** overall. There are reports that did not properly get collected, or companies with **too many missing values** to remain.
- Finding **alternative sources** would likely require a paid subscription but could help with interpretability and accuracy.
- Incorporating **NLP models** which take in text from the annual and quarterly filings could give additional context.
- The model universe could potentially be **expanded beyond the NYSE**.