Machine Learning Applications in Credit Risk Modeling

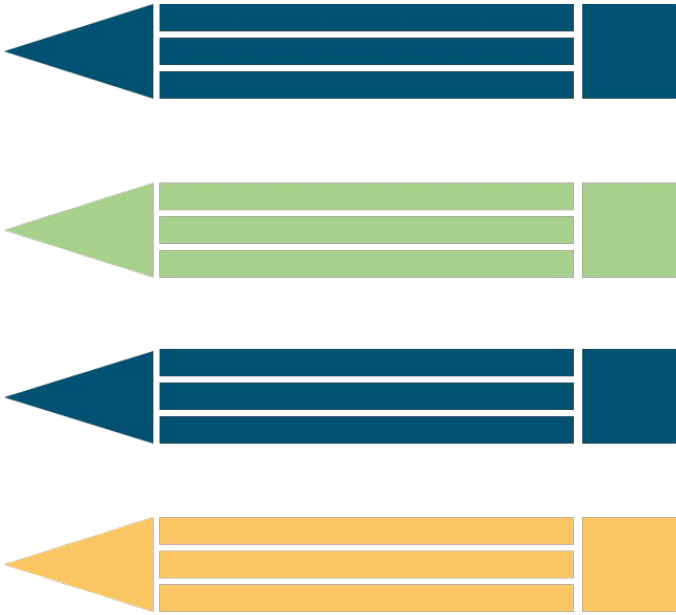
by Yuwei Wang, Viraj Thakkar, Ashwin Siripurapu and Steven Dornberg

October 13th, 2020

Agenda

A structured approach to creating a model

- Business Problem Statement
- Data Description
  - Description of variables
  - Data Preparation
  - Feature Importance
- Model Description
  - Model Introduction
  - Model Evaluation
- Further Discussion



Counterparty Risk Assessment Models

Why Machine Learning?

Business Problem Statement: How can we effectively and efficiency measure Counterparty Risk utilizing a risk model?

Factors and goals to consider in Counterparty Risk Modeling

- Thoroughly evaluate a company's financial metrics to determine variable importance
- Outputting a default risk decision based on systematically analyzing a combination of multiple variables
- Additionally, sometimes we rather know the probability of a default, rather than yes/no, to stratify risk

Benefits of Machine Learning in Evaluating Credit Risk

- ML models can determine which financial metrics weigh most heavily in determining default risk in the dataset
- Machine Learning models hold a distinct advantage over if-else models (static deterministic environment)
- Utilizing past data to assign weights to financial metrics, we can output a probability likelihood that a company defaults

Dataset Description & Features

The data that drives the machine learning model determines its effectiveness

Data Insights for Credit Risk Modeling

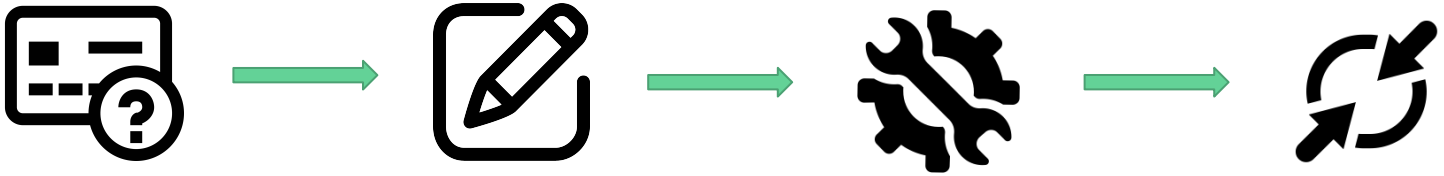
- In order to make future predictions, a machine learning model requires a past dataset to learn from
- Models are only as good as the data they learn from, regardless of the algorithm
- Can the magnitude of a company's financial assets and liabilities, as well as their YoY change, help predict risk?

Dataset Features

- Contains ~4,200 past samples of a company's financial metrics, and whether they defaulted
- Various metrics (features), such as yearly data (ie 2016 TDE), as well as Year-over-Year metrics
- Additional features that were engineered as a result of data preparation and imputing missing values

Data Preparation

Preprocessing steps of data to enter machine learning models



Create missing indicators → Fill in missing values → Fix questionable variables → Standardization

- **Create missing indicators**: add an indicator for each existing variable, to indicate whether the original value was missing (1) or not (0). The missing indicators will be included in model training as features
- **Fill in missing values**: Since most variables are highly skewed, we fill missing values with medians of each variable, rather than mean
- **Fix questionable variables**: All yoy variables are obviously wrongly calculated (negative sign missing and decimals misplaced), so we recalculated yoy variables defined by var_17 - var_16
- **Standardization**: Scaled all variables to [0,1], to fit the requirements of certain models (for example, logistic regression)

ML Models Implemented

Generating novel statistical insights from data

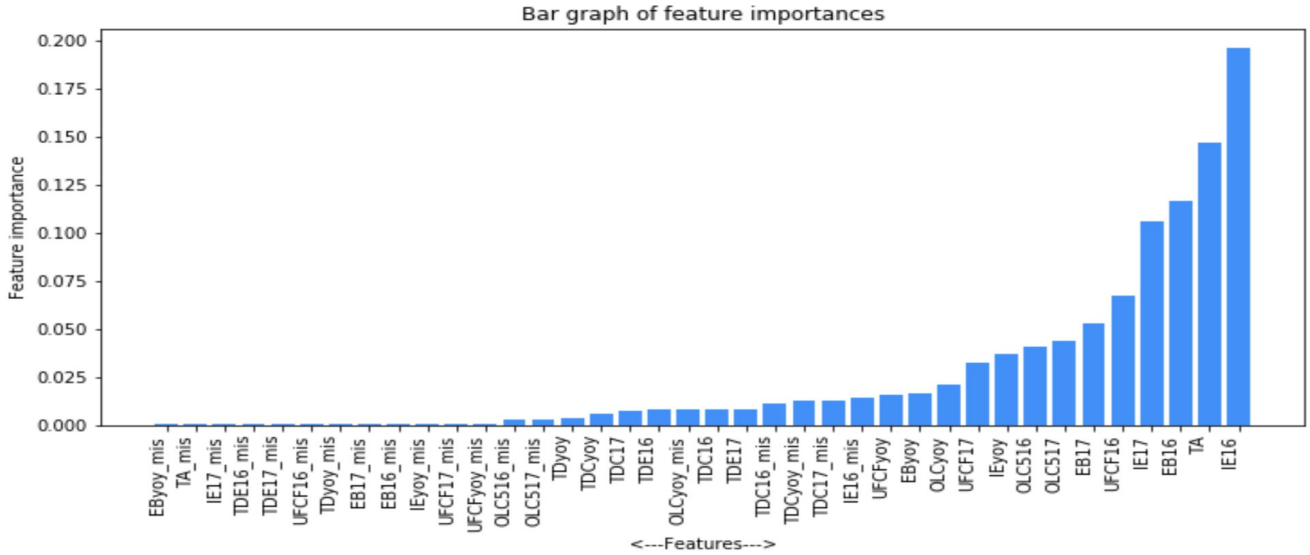
Classification models

- Logistic Regression
- Decision Tree
- Random Forest (best performance)
- Gradient-Boosted Random Forest

Feature Importance

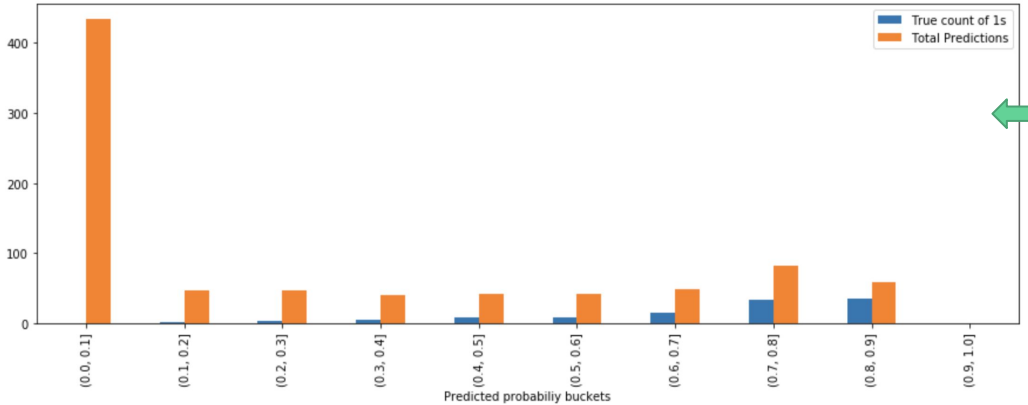
Gauging the most predictive features

- Interest expenses dominate
- Debt-to-capital and debt-to-equity less important than total assets

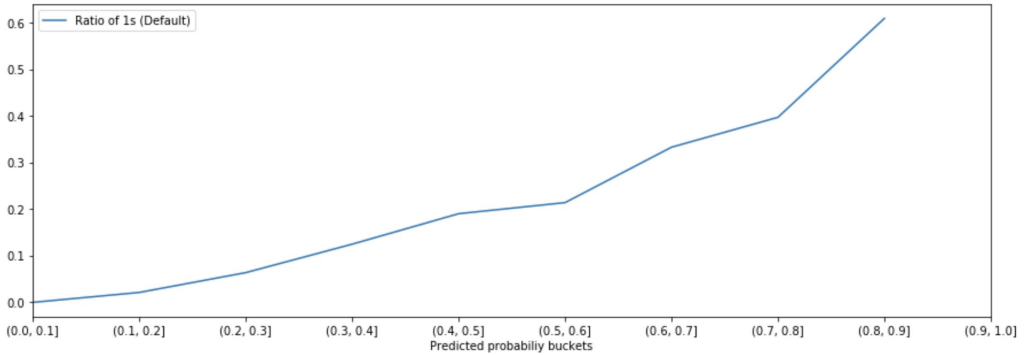


Predictions and Probabilistic Insights

Analyzing outcomes of probabilistic models



Predicted probabilities of Default by RF model

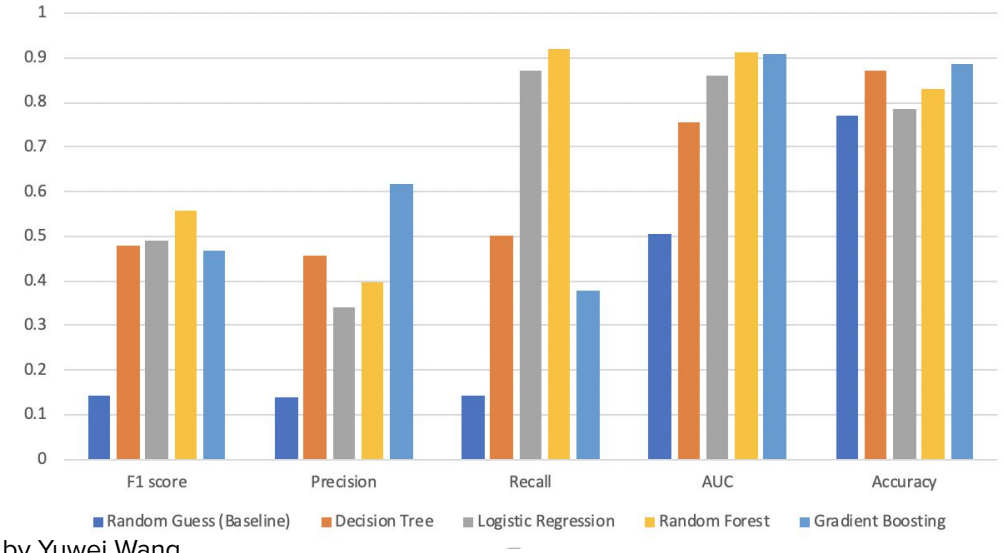


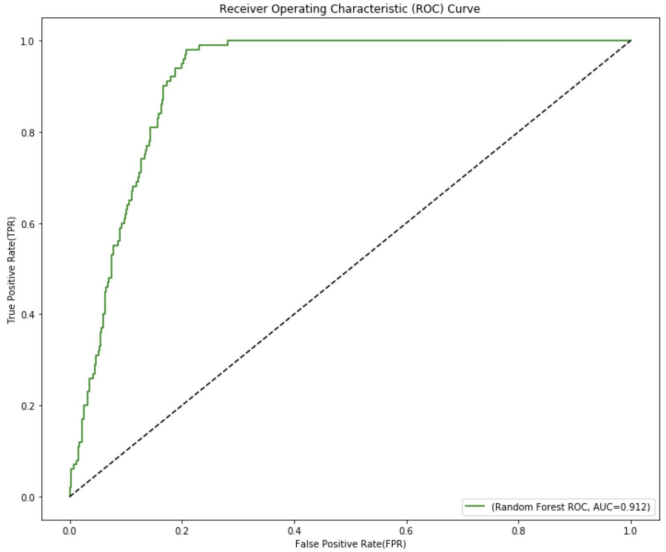
Model Evaluation

Which model to choose?

- **Precision**: proportion of real positive among all predicted positive
- **Recall**: proportion of predicted positive among all real positive
- **F-1 score**: harmonic mean of precision and recall
- **Accuracy**: proportion of correctly assigned observations
- **AUC**: area under receiver operating characteristics curve

	F1 score	Precision	Recall	AUC	Accuracy
Random Guess (Baseline)	0.142	0.139	0.144	0.504	0.769
Decision Tree	0.478	0.456	0.5	0.757	0.87
Logistic Regression	0.489	0.339	0.87	0.862	0.784
Random Forest	0.556	0.398	0.92	0.912	0.83
Gradient Boosting	0.469	0.617	0.378	0.91	0.887





Further Work

Paths forward and areas of improvement

Data Improvements

- More and better data
  - Company size, growth, history data
  - Sectors, industries, and exposure to global macro events
- Longer history to train on

Modeling Considerations

- How frequently to re-train?
- On what time horizon should we predict default?