Predicting Financial Sector Category & Risk

By: Coleton, Jimeshkumar, Van

Project Overview

We have gathered data on publicly traded companies; some operate in the Financial Sector, others do not. The purpose of collecting this company data is to create a Machine Learning mechanism that can correctly classify which companies are in the financial sector and which aren't (binary, 0 or 1). Then, another Machine Learning mechanism will be created to determine a financial risk score for the companies within the financial sector (scale, 0 to 1).

This will help us (The Fed) to maintain the overall stability of the financial system, including the regulation of key non-bank financial institutions.

Table of Contents

Part 1: Exploratory Data Analysis (EDA)

EDA

- The dataset we are working in contains 73 different features across about 500 companies.
- These features contain a range of values, most of which pertain to company financial statistics, along with other character based values such as the description of the company.
- Feature Examples Include:
  - VWAP: Volume Weighted Average Price of Company Stock
  - exchangeCountry: Country where the company's stock is listed
  - Name: The name of the company
  - Operating Margin: Operating income divided by net sales
  - Short Term Debt: Debt that is due within one year

EDA

- The dataset has been split into training and testing sets, which will allow us to create machine learning algorithms to predict our target features - Financial Sector & Financial Risk
- Training Set contains 415 companies across all 73 features
- Testing Set contains 104 companies across 71 features (The target features are not included in the testing set, as those are what we are trying to predict)

```
training set shape: (415, 73) testing set shape: (104, 71)
```

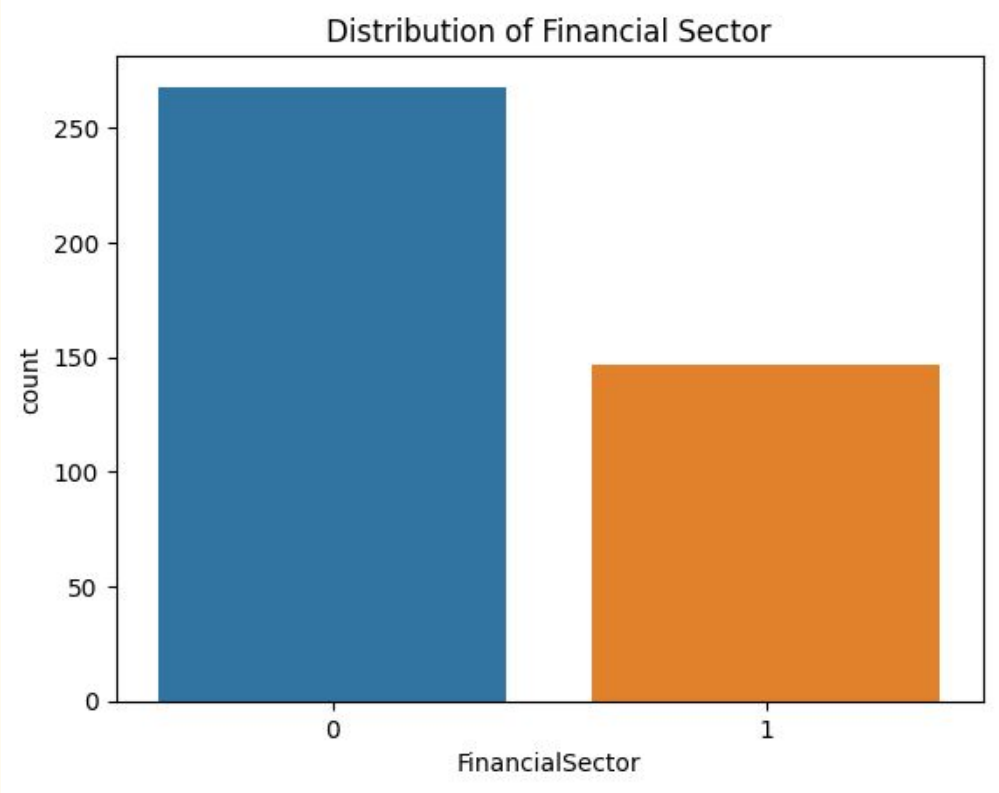
EDA

Financial Risk Description

```
# Display basic statistics of 'FinancialRisk'
print(df_train['FinancialRisk'].describe())
```

```
count    415.000000
mean       0.049077
std        0.117747
min       -0.102909
25%        0.000000
50%        0.000000
75%        0.060544
max        0.998330
Name: FinancialRisk, dtype: float64
```

Financial Sector Distribution Chart



Part 2:
Feature Selection

Feature Selection for Classification

The classification of whether a company is in the Financial Sector or not will be binary.

- A value of 1 means the company is in the financial sector
- A value of 0 means the company is not in the financial sector

Before selecting a classification model, we first had to select the features that we felt would be the most influential in determining whether a company is in the financial sector or not.

The primary feature we chose to use was **businessDescription**, as this was chosen with basic human inferentials as the business description variable describes what exactly the company does.

Feature Selection for Classification

For the rest of the variable selection, we ran a RandomForestRegressor analysis of the Financial Sector against all of the features in the dataset, leaving us with a list of the 15 most important (highly correlated) variables to Financial Sector.

- It is important to note that this was done in an effort to condense feature use in our models to avoid overfitting problems
- We decided to keep variables with a score equal to or greater than .100, leaving us with 5 total variables (including business description).

	Feature	Importance
0	B/P	0.014371
1	CF/P	0.016333
2	Capital Expenditure	0.040841
3	Depreciation	0.108493
4	Earnings Growth (3Y)	0.022311
5	Earnings Growth (5Y)	0.012899
6	FCF/P	0.011037
7	Long Liabilities	0.362171
8	Operating Expense	0.042569
9	R&D	0.006208
10	SG&A	0.117956
11	Sales	0.018740
12	Sales Growth (4Y)	0.013936
13	Short Term Debt	0.014845
14	Working Capital	0.197290

Feature Selection for Classification

The 5 features we selected for our classification model are:

businessDescription: A brief description of the company's business.

Depreciation: The allocation of the cost of assets over time.

Long Liabilities: Long-term financial obligations.

SG&A: Selling, General, and Administrative expenses.

Working Capital: Capital used in day-to-day operations.

Feature Selection for Regression

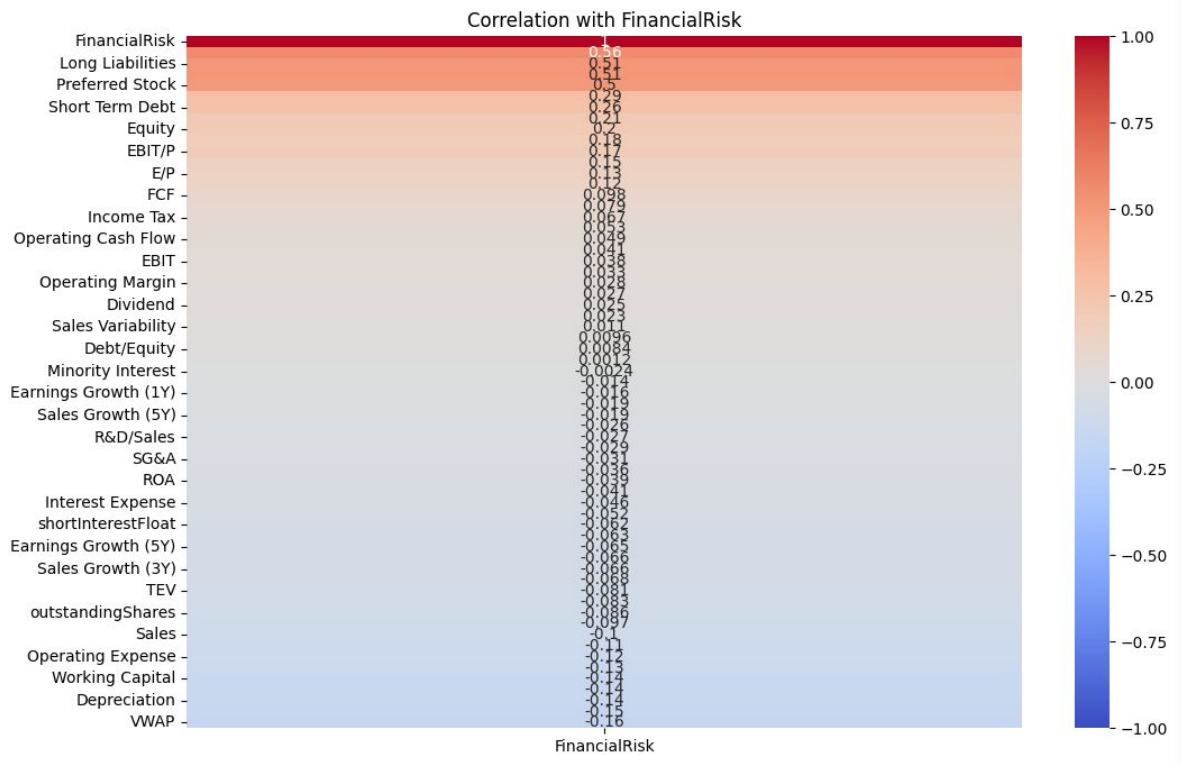
The regression analysis of a company's Financial Risk score will be a continuous value between 0 and 1

- A value of 1 means the company is very financially risky
- A value of 0 means the company is not in the Financial Sector, and should not be assessed.

The process of determining the best features to use in our regression model is very similar to the process in which we followed for the classification model.

This time, we decided not to automatically include business description, and determine which variables to use solely through our feature importance testing

Feature Selection for Regression



We first created a correlation heat map for the numeric values which would show the features most correlated with Financial Risk. (This was through the base .corr() function in python)

Feature Selection for Regression

Next we ran another RandomForestRegressor analysis - this time of Financial Risk against all of the features in the dataset, again leaving us with a list of the 15 most important (highly correlated) variables to Financial Risk.

- We again decided to keep variables with a score equal to or greater than .100, leaving us with 5 total variables (including business description). But there were only 2
- So we decided to also use 3 other features that were seen with high correlation on the heat map

	Feature	Importance
0	VWAP	0.340014
1	B/P	0.013030
2	CF/P	0.015685
3	Capital Expenditure	0.025541
4	Cash	0.015220
5	Depreciation	0.048214
6	E/P	0.021737
7	EBIT/P	0.013120
8	EBIT/TEV	0.030090
9	Equity	0.033188
10	FCF/P	0.019405
11	Long Liabilities	0.332712
12	R&D	0.000113
13	Short Term Debt	0.072800
14	Working Capital	0.019129

Feature Selection for Regression

The 5 features we selected for our regression model are:

VWAP: Volume Weighted Average Price of the company's stock.

Depreciation: The allocation of the cost of assets over time.

Long Liabilities: Long-term financial obligations.

Short Term Debt: Debt that is due within one year.

Preferred Stock: Stock with priority over common stock in dividend payment.

Part 3:
Model Selection

Along with
-Hyper Parameter Tuning

Model Selection for Classification

For our classification model, we first had to ensure that the model would be able to interpret the character values we had brought in with the business description variable.

- This was taken care of by using a vectorizor that would, simply stated, break down the structure of the business descriptions and assign numerical scores

```
from sklearn.feature_extraction.text import TfidfVectorizer

# Text feature processing
vectorizer = TfidfVectorizer(max_features=5000, stop_words='english')
X_train_text = vectorizer.fit_transform(X_train['businessDescription'])
X_test_text = vectorizer.transform(X_test['businessDescription'])
```

Model Selection for Classification

The first model we used for the Classification analysis was a RandomForestClassifier. We decided to use a random forest classifier for several reasons, but mainly because:

- It is an ensemble learning method, meaning it builds multiple decision trees during training and merges them together to get a more accurate and stable prediction.
- It has high accuracy in classification tasks. They are robust and perform well on both small and large datasets.
- It has no need for Feature Scaling as it can handle both categorical and numerical features, making them convenient for datasets with a mix of data types.

After running a RandomForestClassifier Model, we returned an accuracy score of .96, which is very good.

```
# Validate the model
val_predictions = clf.predict(X_val_split)
accuracy = accuracy_score(y_val_split, val_predictions)
print(f'Validation Accuracy: {accuracy}')
```

```
Validation Accuracy: 0.963855421686747
```

Model Selection for Classification - Hyperparameter Tuning

In order to make the model even better, we bagan hyperparameter tuning. We created a GridSearch to search for the best parameters that would benefit our model.

- We used the gridsearch because it automates the process of finding the optimal hyperparameters, saving time and effort compared to manual tuning. It provides an exhaustive search over a specified hyperparameter space, ensuring that the selected hyperparameters lead to a well-performing model.
- We then applied the best hyperparameter values to our model, giving us terrific results.

```
Accuracy: 1.0000

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        59
           1       1.00      1.00      1.00        45

    accuracy                           1.00       104
   macro avg       1.00      1.00      1.00       104
weighted avg       1.00      1.00      1.00       104
```

All of the 1.00's tell us that the model is nearly perfect, which is great.

Model Selection for Regression

For our regression model, we first decided to try our luck with the random forest again, using a RandomForestRegressor model.

The benefits of this type of model are very similar to that of the RandomForestClassifier model. However, we are not using any text based features in this model, so we may want to find an alternate model.

```
[104 rows x 2 columns]
Mean Squared Error on Training Set: 0.0005657044875897339
R-squared on Training Set: 0.9590985364875 06
Cross-validated R-squared scores: [0.79527902 0.77992084 0.78959127 0.6262989  0.63606547]
```

The output seen is very inspiring, with having such a high R-squared as well as a very low MSE on the Training Set.

Let's still try another model.

Model Selection for Regression

The second regression model we tried to use was an XGBoost Model. We wanted to use XGBoost for a variety of reasons, but mainly:

- The high performance - XGBoost is designed for speed and performance. It often outperforms other machine learning algorithms and can handle large datasets efficiently.
- XGBoost includes regularization terms in its objective function, such as L1 (Lasso) and L2 (Ridge) regularization. This helps prevent overfitting and improves the model's generalization to unseen data.
- XGBoost uses a depth-first approach for tree construction and prunes trees during the building process. This helps control the complexity of the model and improves its generalization.

```
Mean Squared Error on Training Set: 1.3713937957166785e-06
R-squared on Training Set: 0.9999008457338994
Cross-validated R-squared scores: [0.81085732 0.75257687 0.28209983 0.46983295 0.64070443]
```

The feedback here is very positive. Let’s continue with tuning the hyperparameters on this model

Model Selection for Regression - Hyperparameter Tuning

In order to make the model even better, we bagan hyperparameter tuning. We created a GridSearch on this model as well to search for the best parameters that would benefit our model.

Surprisingly, the model seemed to become slightly worse, but still incredibly good.

```
Mean Squared Error on Training Set: 0.00025509227519958697
R-squared on Training Set: 0.981556364470699
Cross-validated R-squared scores: [0.84940408 0.78920616 0.54097938 0.64879583 0.72453188]
```

Part 4:
Final Model

Final Model - Classification

Our Final Model for the Classification of Financial Sector is the RandomForestClassifier Model with HyperParameter tuning applied using GridSearch to find the best parameters.

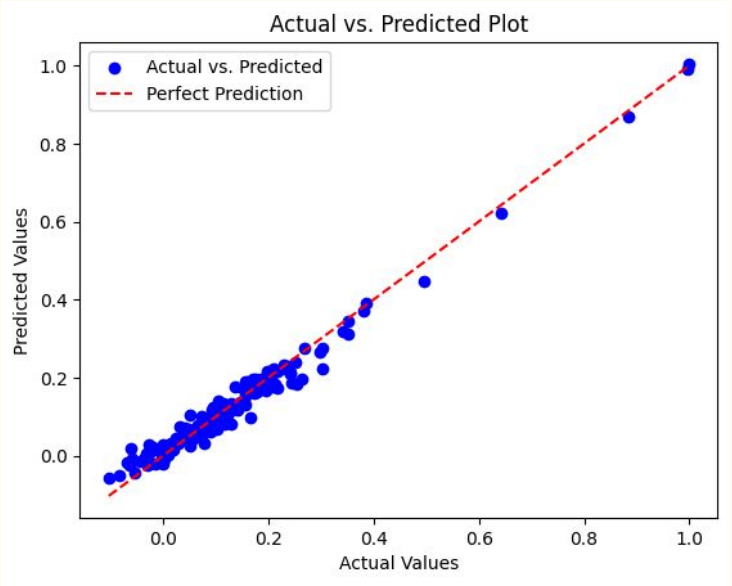
This model returned a near perfect accuracy score, as the F-1, Recall, and Precision Scores were all equal to 1.00, leading the overall accuracy to be 1.00. Now, the model probably isn't entirely perfect. However, we can say with confidence it is very close to it.

Part of the model's success is due to the strategic use of select features.

Final Model - Regression

Our Final Model for the Regression of Financial Risk Score is the XGBoost Model with HyperParameter tuning applied using GridSearch to find the best parameters.

Even though the Rsqaured and MSE were not as good as the XGBoost with no Hyperparameter tuning, the cross validated Rsquared scores were not only more consistent with each other, but also better scores in general. This left us to believe that the model with the hyperparameter tuning is actually more reliable than that without it.



Part 5:
Feature Importance & Explanation

Feature Importance & Explanation - Classification

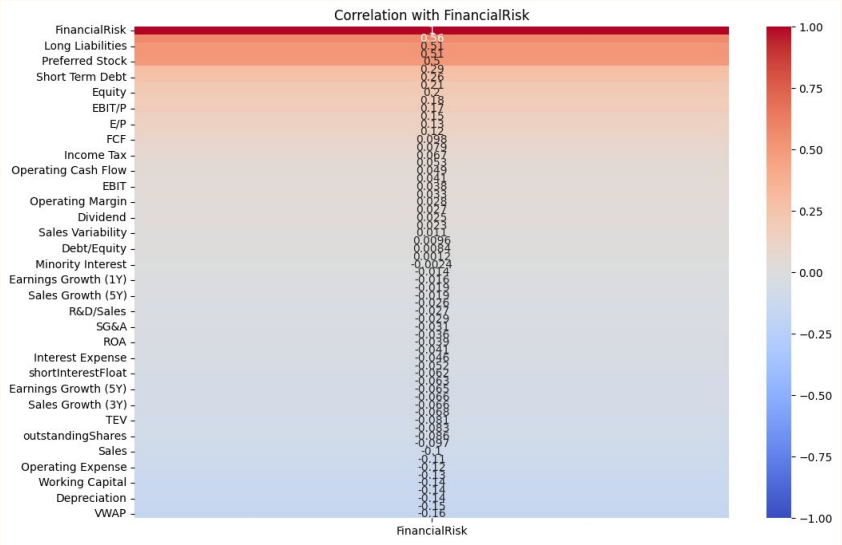
As mentioned previously, we came to the conclusion that the most important feature in determining the classification of a company's financial sector is the business description - throughout our process we did make a few models without the categorical value of business description and none were as good as our current model.

It was no surprise to see long liabilities hold as much importance as it did - companies within the financial sector often do have long term financial obligations.

Feature Importance & Explanation - Regression

- Comes as no surprise for FinancialSector to hold as much importance as it does
- VWAP makes sense to have the second most weight - as the volume of the company stock increases with price, the financial risk in the stock decreases (this was seen in negative correlation in heat map shown earlier)

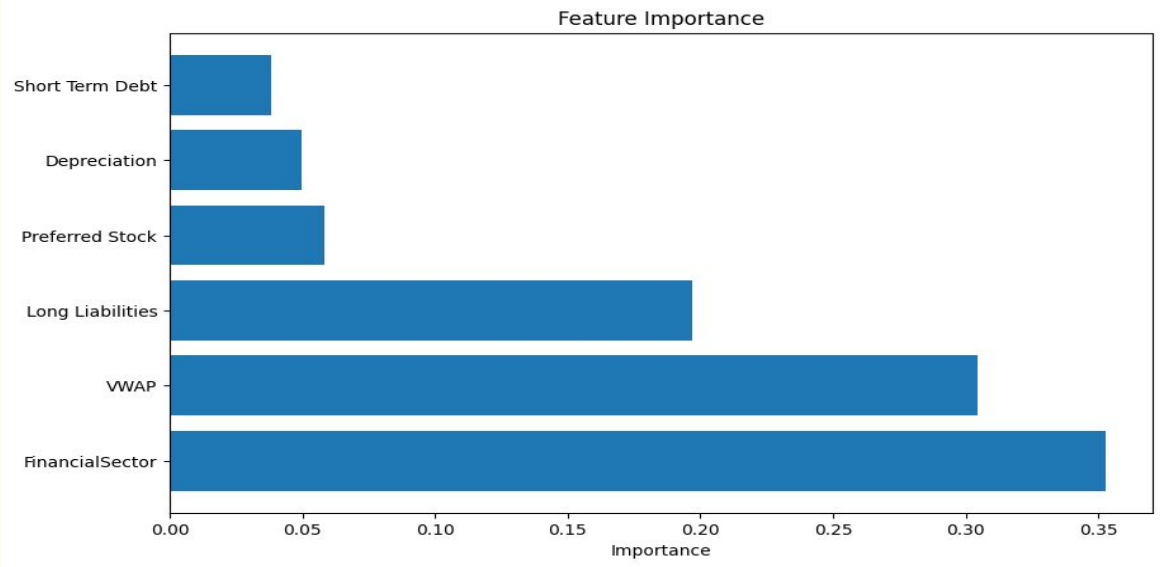
	Feature	Importance
0	FinancialSector	0.352798
1	VWAP	0.304323
2	Long Liabilities	0.196776
4	Preferred Stock	0.058357
5	Depreciation	0.049671



Heat Map Again for Pos/Neg Relationship Guidance

Feature Importance & Explanation - Regression

- Long liabilities pose financial risk, which also makes sense
- Preferred stock and short term debt have similar importance and are positively correlated with Financial Risk
- Depreciation has a negative correlation and holds a similar importance to the surrounding features



Part 6: Conclusion

Conclusion

- BusinessDescription and Long Liabilities helped in aiding the classification of companies within the financial sector.
- Our preferred regression model was an XGBoost Model that used gridsearch to find the best parameters.
- In order to avoid further financial crises, the fed should refer to our model, as well as keep a keen eye on companies within the financial sector that have a lower VWAP with a larger amount of long term liabilities.

Thank you!