COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

## Capstone Report - II

# Business Health Simulator with Capital One

Industrial affiliate : Peter Deng, Chaitanya Dara

Faculty advisor: Adam Kelleher


**Made by:**

Ankita Agrawal, UNI: aa4229

Akanksha Rajput, UNI: ar3879

Gurkanwar Singh, UNI: gs3006

Sheetal Reddy, UNI: kr2793

# Contents


**1 Data Collection & Visualization** **2**
- 1.1 Data Collection Steps: . . . 2
- 1.2 Features Included: . . . 2
- 1.3 Preprocessing EBITDA (old target) for new data range: . . . 2
- 1.4 Change in target variable: . . . 3
- 1.5 Visualization of new target variables: . . . 4

**2 Final Data Preprocessing & Cleaning** **4**
- 2.1 Missing value techniques: . . . 4
- 2.2 Lagging: . . . 4
- 2.3 Model selection based on target: . . . 5
- 2.4 Micro vs Macro inspection of Data: . . . 5

**3 Models and Predictors** **5**
- 3.1 Classification Techniques: . . . 5
- 3.2 Regression Techniques: . . . 6
- 3.3 LSTM: . . . 6
- 3.4 Bayesian Model: . . . 7
- 3.5 Time Series: . . . 7

**4 Results:** **7**
- 4.1 Classification Results . . . 7
  - 4.1.1 Missing Value Imputation Techniques . . . 7
  - 4.1.2 Lagging Techniques . . . 8
- 4.2 Regression results . . . 8
- 4.3 LSTM Results . . . 9
- 4.4 Bayesian Results: . . . 9
- 4.5 Time Series Results . . . 9
- 4.6 Final Results: . . . 10
  - 4.6.1 Feature Importance . . . 10

**5 Report Contribution** **10**

**6 Acknowledgement** **11**

**7 References** **11**

# 1 Data Collection & Visualization

The data collection in phase-I was much smaller and was only for the years 2018-2019. The aim there was to familiarize ourselves with the problem statement better, learn the correlation, distributions and patterns which exist in the data set and also thoroughly understand the contribution of each feature in the context of the problem posed. In the second phase we aim to dig deeper by using a larger data set. Total data gathered by us is for more than 10 years [2008-2019].

## 1.1 Data Collection Steps:

- Dataset for a 10 year period is huge and hence a lot of effort went in learning VBA to be able to write macros which can automate the process of data collection using bloomberg terminal.
- 1400 small business tickers as recognized by Vanguard were scraped from this website using a selenium based python script. This was then fed into our macros to extract the larger data at once.

## 1.2 Features Included:

We have included macroeconomic indicators to our data along with other features that were outlined in the first report.

| **Features** | **Macroeconomic Features** | **Target** |
|---|---|---|
| Price to Earnings Ratio | Un-employement Rate | Free Cash flow (CF) |
| Price Cash Flow Ratio | CPI | Operating Margin (OM) |
| Price Book Ratio | GDP CURRY Idx | Profit Margin (PM) |
| Dividend yield | Financial Stress Idx | EBITDA (dropped) |
| Market Valuation | Consumer Confidence Idx | |
| | Jobless Claims | |
| | ISM Manufactoring Idx | |

Table 1: Features and Target in the dataset

The target variable 'EBITDA' was used in the phase-I, which worked fine with smaller data and gave reasonable results on our baseline model. We performed the following steps on larger data set to preprocess the target EBITDA.

## 1.3 Preprocessing EBITDA (old target) for new data range:

- Several companies had missing EBITDA values. The values that were missing in the beginning or end of the timestamp could be explained by the company not being established. Few companies had missing values in the middle of the dataset that were difficult to explain.

- Our problem is modeled in a way that features from present quarters are used to predict target in future (say for next 1Q,3Q or a year).
  This lagging effect is included in our training data, and whole target data is shifted by the lagging value (1Q,3Q or 1year). Under such case, missing values in any one of the quarter starts impacting more than 1 quarters and results in manual handling of lot many rows than expected.
- To handle these missing EBITDA values we explored the following approaches:
  - Companies with more than 75% EBITDA values (target rows) missing were dropped.
  - Or, rows with EBITDA value missing in the current quarter was skipped along with rows in the previous quarter.
  - Or EBITDA values were imputed using techniques such as Random Forest.

The binary classification model tried using the percentage change values as either positive and negative class performed poorly at about 50% accuracy which is as good as a random prediction. Thus, instead of imputing the target variable and handling it manually, a decision was made to drop it and consider other variables.

### 1.4 Change in target variable:

The lack of clear data for this EBITDA forced us to change how we perceive the health of a business. Now, to define the health of a business we have finalized three variables instead of one:

- **Free Cash Flow:** Free cash flow (FCF) is a measure of how much cash a business generates after accounting for capital expenditures such as buildings or equipment. This cash can be used for expansion, dividends, reducing debt, or other purposes.
- **Operating Margin:** The operating margin measures how much profit a company makes on a dollar of sales, after paying for variable costs of production, such as wages and raw materials, but before paying interest or tax.
- **Profit Margin:** Profit margin is one of the commonly used profitability models to gauge the degree to which a company or a business activity makes money. It represents what percentage of sales has turned into profits.

Further we have expanded our feature set by adding macroeconomic indicators such as CPI, GDP, Unemployment rate, etc.

There were a few outliers in the new target variables with extreme values around +/-700000 which was not at all intuitive as our new target variables are all percentage values. Thus, our target variables are now bounded between -100 to 100 by dropping those rows with outliers.

### 1.5 Visualization of new target variables:

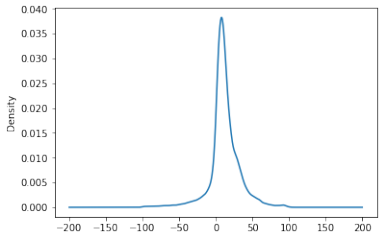


Figure 1: Free Cash Flow

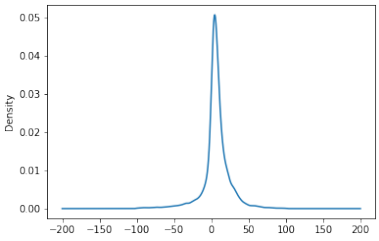


Figure 2: Profit Margin

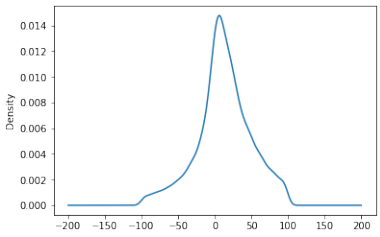


Figure 3: Operating Margin

## 2 Final Data Preprocessing & Cleaning

### 2.1 Missing value techniques:

There are many missing values in our features that need to be processed before being fed into any model. One way to handle these missing values is to impute them. We explored the following imputation techniques:

- **Random Forest Imputation:**
  This technique refers to the imputation of missing values using Random Forests. It is an iterative method of imputing values. For continuous predictors, the imputed value is the weighted average of the non-missing observations, where the weights are the proximities. For categorical predictors, the imputed value is the category with the largest average proximity.
- **Mean Imputation:**
  This is an imputation method used to fill in missing continuous NA valued. The NA value is replaced by the mean of all the values present in that feature.
- **Median Imputation:**
  This is an imputation method used to fill in missing continuous NA valued. The NA value is replaced by the median of all the values present in that feature.
- **Forward fill/Backward fill:**
  An imputation method used for both continuous and categorical variables where the values that are missing/are NA are replaced by the value succeeding or preceding it.

### 2.2 Lagging:

The aim of the project is to help predict the health of a business in the future. To make sure our model is doing so, we shift our target variable up by a few rows. This ensures that the features of the present time are paired with a future target value. So when we input the features of the current time, the model is able to predict the target for a future time step.
We have used lag of the following values for our predictions:

- Shifted by one quarter (Lagging 1)
- Shifted by two quarters (Lagging 2)
- Shifted by three quarters (Lagging 3)
- Shifted by a year (Lagging 4)

### 2.3 Model selection based on target:

All of our target variables are continuous fields with varying ranges.

- Regression techniques can be used to predict future values.
- Classification techniques: To simplify the problem of predicting business health, we have modeled our problem into one of classification. The approach used was to consider all positive values as growth in business health and all negative values as decline in business health.

The two ways classifications were implemented:

- Binary classification: Positive values in target are assigned to 1 and negative are assigned to 0.
- Multiclass classification: The target variable was converted into 4 classes instead of 2. Rows with values <-50 were labelled '-2' , those between (-50,0) were labelled '-1' , those between (0,50) as '+1' and >50 as '+2'.

### 2.4 Micro vs Macro inspection of Data:

The data that we are working with is an amalgamation of 1400 small businesses. There are different levels at which we can look at the data. We have used the following levels for predictions:

- **Company level:**
  We include the 'Company' as a feature and one hot encode it to ensure that the model learns at the company level.
- **Industry level:**
  'Company' column is now replaced with a industry-type column. The classification of these companies into the industry is obtained from SEC website based on the filing sector that they come under.
- **Without any segmentation:**
  We excluded both the company and industry level column to look at the data as a whole.

## 3 Models and Predictors

### 3.1 Classification Techniques:

Once we converted our problem into a classification problem (Section 2.3 ) we experimented with different classification models. The results for this have been mentioned in section 4.1. To preprocess our features before feeding them into the model, we built a pipeline to scale our continuous variables and one hot encode our categorical variables.

The models used for binary classification:

- Decision Tree Classifier
- Random Forest Classifier
- AdaBoost Classifier
- Gradient Boosting Classifier
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis

The models used for multi-class classification:

- Random Forest
- SVC
- Logistic Regression
- Gradient Boosting

### 3.2 Regression Techniques:

The results for this have been mentioned in section 4.2. To preprocess our features before feeding them into the model, we built a pipeline to scale our continuous variables and one hot encode our categorical variables.
The models used for binary classification:

- Linear Regression
- Decision Tree
- Ridge Regression
- Gradient Boosting
- Ada Boost

For better effeciency target variable was scaled while training the model, thus the predicted results were also scaled. These scaled predictions were mapped back to its original domain using mean and standard deviations from the training set.

Also removal of outliers helped the model to learn a lot better.

We have used cross validation to evaluate all these methods. We also tuned the hyperparameters by using RandomizedSearch and GridSearch to test all combinations of hyperparameters.

Along with the hyperparameter tuning we also changed our datasets to experiment with different targets as described in (1.1), different imputation as described in (2.1), different lagging as described in (2.2) and different levels of data as in (2.4)

### 3.3 LSTM:

LSTM is a neural network architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. Data needs to be pre processed before providing it as input into the network. The categorical variable is one hot encoded. The data frame is converted into a matrix of numpy values. The continuous variables are standardized. A time step of one is used as we used 1 quarter lagging i.e shifting target variable by 1 quarter. To tune the parameters better we tried methods of regularization techniques such as dropout.

### 3.4 Bayesian Model:

Bayesian approach allows us to better understand the uncertainty of our model. We used a probabilistic programming language STAN using R (rstan package), to build the Bayesian classifier.

$$y \propto bernoulli\_logit(\alpha + X * \beta)$$

Usually, for the Bayesian approach, we must choose a prior distribution which represents our initial beliefs about the estimate.The posterior distribution is a weighted combination of prior and likelihood.

**Prior of alpha used:** $alpha \propto cauchy\_logit(0, 10)$

**Two types of prior of beta used:**

$$beta \propto normal(0, 1)$$
$$beta \propto cauchy(0, 2.5)$$

The model built using Stan results a set of posterior simulations of the parameters in the model (or a point estimate, if Stan is set to optimize). With more data, the weights shift even more to the likelihood, eventually making prior inconsequential .

### 3.5 Time Series:

Since our dataset contains a sequence of dates for each company, we experimented with time series modeling using ARIMA for a few companies. Autoregressive Integrated Moving Averages (ARIMA) is a technique in which the model is fitted to time series data to predict future points in the series.

We looked at one company Alcoa Corporation (ticker: AA) and found the optimum values of p, d, q and P, D, Q (autoregressive, differencing, and moving average terms ) by fitting the SARIMA model on the Free Cash Flow target variable with combinations of range of values of p, d, q from 0 to 2 and selecting the model with the minimum AIC.

## 4 Results:

### 4.1 Classification Results

#### 4.1.1 Missing Value Imputation Techniques

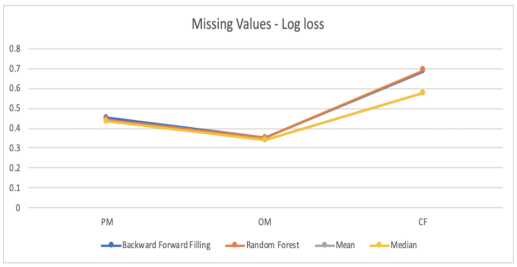


Figure 4: Log Loss

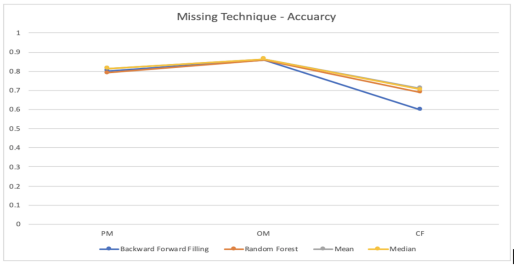


Figure 5: Accuracy

As seen in the above accuracy graphs for the various missing techniques we observe that for all three targets the imputation of mean gives us the best results. Thus, we chose to impute all of our data with mean for the further experiments.

### 4.1.2 Lagging Techniques

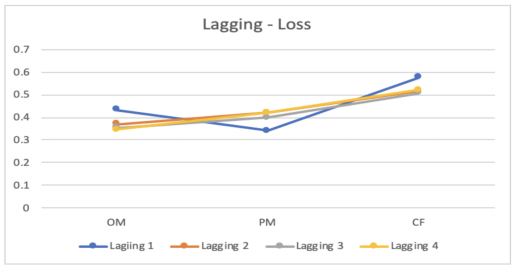


Figure 6: Log Loss

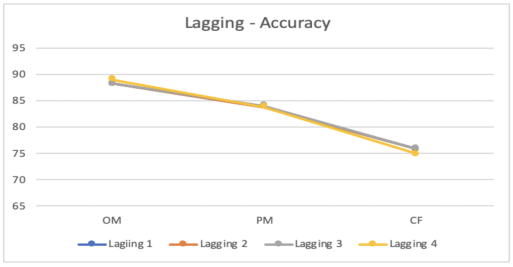


Figure 7: Accuracy

## 4.2 Regression results

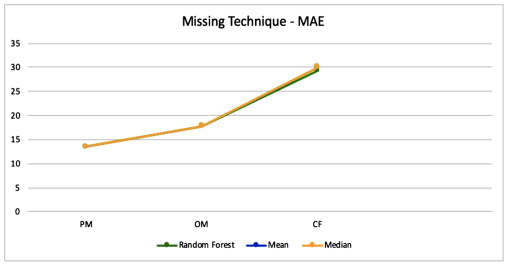


Figure 8: Missing Value MAE

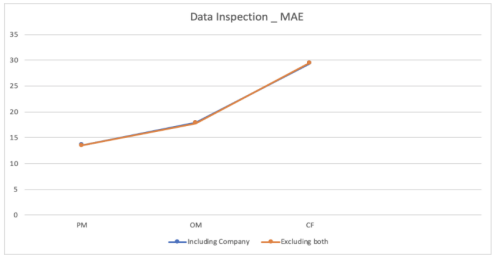


Figure 9: Data Inspection MAE

All the imputation techniques work very similar for regression models, for further analysis we use data that was imputed with Random Forest.
Both the data inspection techniques have similar results, for further analysis we use data that doesn't have company level distinction.

| Target Variable | PM | OM | CF |
|---|---|---|---|
| Linear Regressor | 113.76 | 243.04 | 461.44 |
| Gradient Boosting Regressor | 13.59 | 17.88 | 30.13 |
| Decision Tree | 13.50 | 17.79 | 30.02 |
| Ada Boost | 13.61 | 17.86 | 30.12 |
| Ridge Regressor | 13.60 | 17.75 | 30.10 |

Table 2: After parameter tuning the regression model MAE results are:

### 4.3 LSTM Results

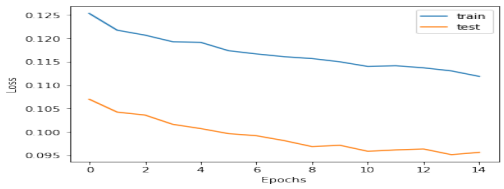


Figure 10: MAE Loss

Above is a graph for LSTM prediction loss values, that utilises Adam optimiser and Dropout of 0.5. As the test result is only from the quarters of 2019, the variance in the data might be less because of which we are achieving test error less than training.

### 4.4 Bayesian Results:

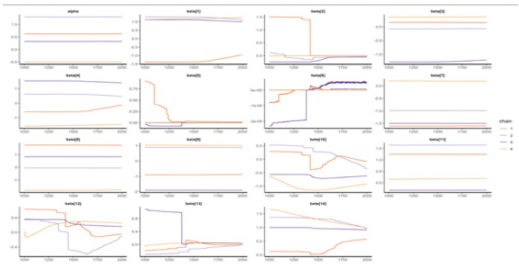

Figure 11: Trace plot

In traceplot, each color represents different markov chain. Since the chains are not mixing well, it indicates that chains are not eventually converging to the stationary distribution i.e. target distribution. Hence results of this model is more likely to be unreliable.

### 4.5 Time Series Results

For the company Alcao Corporation (AA), we fit the SARIMA(1,1,1)x(1,1,0,12) model on the Free Cash Flow target variable and got an MSE of 219842. We obtained the following plot for forecast period Sept 2016 - Dec 2019: and forecast period Sept 2019 - Dec 2023:

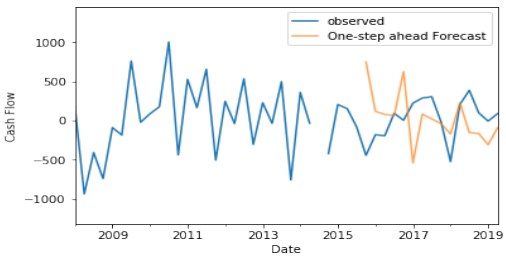


Figure 12: Prediction Vs Actual

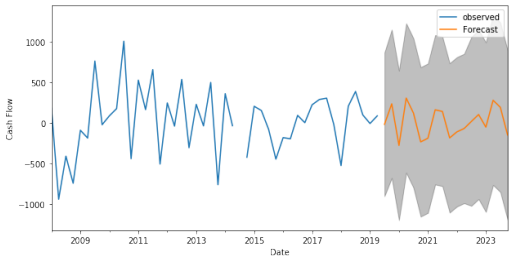


Figure 13: Forecast

### 4.6 Final Results:

We observed that Linear Discriminant Classifier performed with 88% accuracy and Decision Tree Regressor performed with 13.05 MAE on a target with mean std deviation of 37.4. These are the best results obtained so far after parameter tuning.

#### 4.6.1 Feature Importance

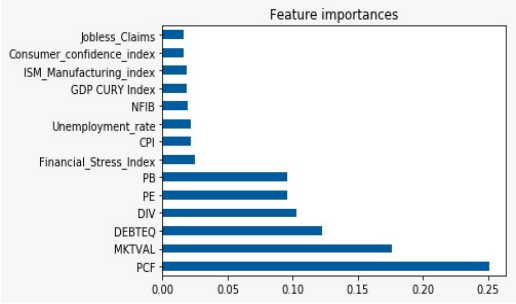


Figure 14: Feature Importance

We can see that financial variables, such as Price-cash flow ratio, which are provided at the company level are more important than macroeconomic features which are the same for all companies.

## 5 Report Contribution

- Akanksha Rajput - Missing Value Imputation,Data merging,Scraping, Bayesian inference techniques (STAN)
- Ankita Agrawal - Preprocessing, LSTM, Regression, Report writing and latex compilation
- Sheetal Reddy - Preprocessing, Classification, Report writing, Model evaluation and comparison, Visualizations
- Gurkanwar Singh - Data collection, Time Series analysis, Results and summary, feature selection and exploration

## 6 Acknowledgement

We would like to thank our faculty advisor Adam Kelleher for his valuable feed-backs and guidance through all the phase of this work. We would also like to extend our gratitude to the team at Capital One (Peter Deng and Chaitanya Dara) for giving us this opportunity to work with them and for their continuous support and mentorship.

## 7 References

1. https://mc-stan.org/rstan/index.html
2. https://quickbooks.intuit.com/r/financial-management/5-financial-kpis-gauge-business-health/
3. https://www.fundera.com/blog/profitability-ratio