Analysis of European Financial Firms

Gaurav Gul Lalwani

Jaskirat Singh

Omkar Pandit

Shweta Mane

Sonakshi Garg

Introduction

In recent years, European economy has gone through many ups and downs. From being in bad shape during its debt crises and the incident of Brexit which happened few years ago to becoming the fastest growing economy in the year 2017, a lot has changed. These uncertainties concern upcoming firms about their scope of sustainability in the current market, success of their company or status of their competitors.

Moreover, once a firm enters the market, next thing that comes along with operating is surviving the crests and troughs in the market. Unquestionably, it comes to predicting the occurrence probability of an event and predictive statistical models built on historical data are certainly the best of all approaches to start from. With all these goals in mind, we build multiple statistical models based on real world data to predict the success of a company and then choose the best ones.

"*Lower trade due to reduced integration with EU countries is likely to cost the UK economy far more than is gained from lower contributions to the EU budget*."(Dhingra, Ottaviano, Sampson, & Reenen, n.d.)

Problem Statement

1. What would be the operating profit for a new company entering the European Market:
    a. Gross Profit
    b. Turnover / Operating Revenue

2. What would be the size of a financial firm (large, medium or small) in the European market?

DATA

Our chosen dataset is 'Amadeus' which is made available by WRDS - Wharton Research Data Services which is a platform providing access to various business research databases.

Amadeus is a big data (in tens of terabytes) platform which has information from over thirty countries and has around 150 fields for about thousands of European financial firms. It was filtered to fetch data of financial firms from only a few countries (keeping in mind the project scope and timeline). After filtering, the data still has approximately 400,000 data points with around 150 fields to further filter from.

This data is not simulated or preprocessed. It is rich and raw, free from any alterations of any type.

Source: https://wrds-web.wharton.upenn.edu/wrds/

Response Variable Selection

The prime indicators that define a company's success are turnover and gross profit, so for regression problems, we wanted to be able to predict the turnover and gross profit of a company in the European market. For classification, we stressed on classifying them in a predefined categories of company size.

PREPROCESSING

The downloaded data, being a part of the big data demanded a significant amount of preprocessing to for it to be managed with the available resources at hand. Number of observations and predictors both were very large and required cleaning to build the models that would answer our research questions. This process was done in the following two parts:

MS EXCEL:

a. The data was downloaded as a comma separated file using STATA for better compatibility with Python.
b. Out of approximately 150 fields, 45 were selected based on background knowledge and subject research. (Fields like phone number, CEO name etc. were removed)
c. The redundancy in the data was removed.

PYTHON:

a. Considering the large availability of data, data points with too many missing values were removed.

b. The outliers in the data were also removed.
c. Missing values were replaced with mean of the column data
d. For all the categorical variables, we added dummy variables for computational purposes.
    - The "Company Category" field was manipulated to make four new field that would have binary inputs against each level of measurement (very large, large, medium, small).

SPLITTING OF DATASET

Data was split such that we had 70% observations to train our models and 30% to test them.

Number of training observations ~ 280000

Number of test observations ~ 12000

ASSUMPTIONS

The project builds up on the following assumptions:

1. The data being used is not skewed and comes in its raw form.
2. The analysis given by the models applies to all the European countries while it is built from only the countries which contain the maximum data compared to others.

WEAKNESSES

- Given the time frame for the project, it was not feasible to carry out the further preprocessing of the data.
- In the absence of a server system, there was limited scope to build and test all the statistical models on local machines without cutting down on data, and as a result, much of the data was sacrificed.
- Limited ability of finding interaction terms, as in the scenario of analyzing the real-world data, interaction among the predictors could be a crucial factor.

- Lastly, if there exist solutions in the data which could be potential answers to our problem statements could may turn out as a weakness if there aren't any relations in actual.
- Also, the problem of multicollinearity gave issues while selecting models clearly depicting the cases of over-fitting.

ANALYSIS

INITIAL ANALYSIS:

We tried to run a selection of models on our entire dataset, to understand some of the weaknesses of our data. From the initial analysis we could conclude several drawbacks of using the entire dataset, the results concluded that our models were:

1. Overfitting: - All the models indicated overfitting as instead of low training MSE score the test MSE scores were higher.
2. High variance covered: - Building models on the entire dataset covers almost entire variance of the data and it was done intentionally to understand the attributes better.
3. Higher R-squared value: - The overfitting and high variance coverage were both evident from the high R-squared values that all the models gave.

SELECTION OF VARIABLES:

After the analysis of building our models with all our predictors we faced many challenges with low predictive accuracy. To overcome these challenges, we decided to reduce the dimensionality of our dataset. The processes employed to deduce feature importance from all the available features were as follows: -

1. **Correlation:** - The important factor which drives the selection of the features for our model was the correlation. We used the correlation function to deduce the correlation of all the features with respect to the response. We then observed features with the

highest correlation ratio. We kept a threshold of 0.5 for this correlation and neglected the features below this threshold.

2. **Importance: -** The correlation factor used in the previous step did not take into account the multi collinearity between variables. This problem was taken care of by the importance() function of Random Forest Classifier. When we ran the model previously using all the predictors we ran our model using the Random Forest Classifier and used the importance function to find the impact of different predictors on our response. We were clearly able to see the percent impact of different predictors on the corresponding responses.

➢ **For response variable Gross Profit**

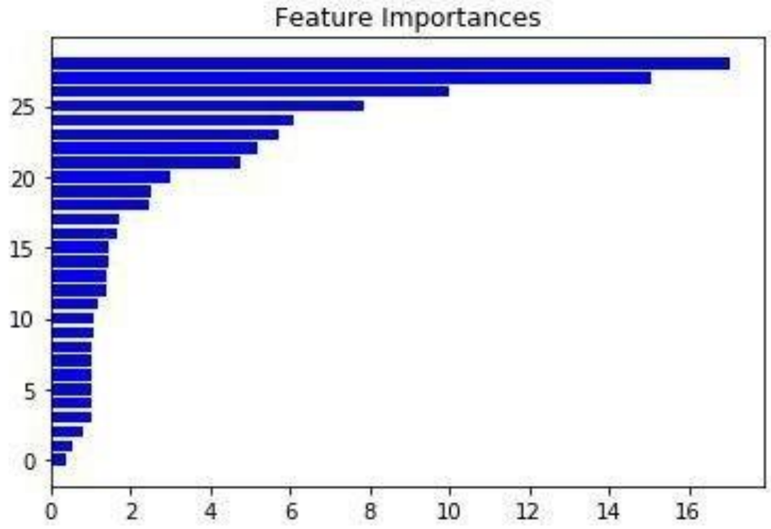


Figure 1: *Output of Importance function for Gross Profit*

Most important predictors for response Gross Profit are as below (in descending order)

•Shareholders funds (17.03%)

- Financial expenses (15.08%)

- Costs of goods sold (10.01%)
- Current Liabilities: loans (7.83%)

***Table 1**: List of all predictors with their percentage of impact on Gross Profit:*

Name of the predictor	Impact on the response (%)
Shareholders funds	17.03
Financial expenses	15.08
Costs of goods sold	10.01
Current Liabilities: loans	7.83
Current assets: stocks	6.09
Other fixed assets	5.73
Current assets: debtors	5.19
Other operating expenses	4.74
Other operating expenses	3.02
Taxation	2.52
Sales	2.47
Total assets	1.73
Other shareholders funds	1.66
Tangible fixed assets	1.46
Current Liabilities: creditors	1.45
Enterprise value	1.41
Current liabilities	1.37
Tangible fixed assets	1.16
Total shareh. funds & liab.	1.08
Costs of goods sold	1.08
Cash & cash equivalent	1.04
Current assets	1.04
Working capital	1.04
Net current assets	1.04
Shareholder funds: capital	1.04
Other fixed assets	1.04
Other current assets	0.99
Financial revenue	0.65

➢ **For response variable Turnover**

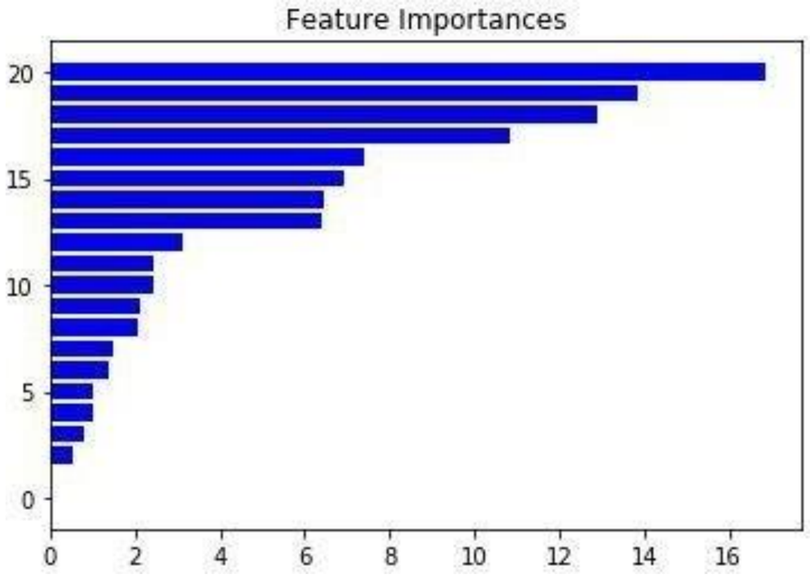


Figure 2: *Output of Importance function for Operating Revenue/Turnover*

Most important predictors for response Operating Revenue/Turnover are as below (in descending order)

- Other operating expenses (16.85%)
- Other fixed assets (13.84%)
- Current Liabilities: creditors (12.91%)
- Costs of goods sold (10.84%)
- Current assets (7.4%)

Table 2: *List of all predictors with their percentage of impact on Operating Revenue/Turnover:*

Name of the predictor	Impact on the response (%)
Other operating expenses	16.85
Other fixed assets	13.84
Current Liabilities: creditors	12.91
Costs of goods sold	10.84
Current assets	7.4

Total shareh. funds & liab.	6.92
Cash & cash equivalent	6.47
Current assets: stocks	6.38
Other current assets	3.13
Taxation	2.45
Current assets: debtors	2.41
Tangible fixed assets	2.1
Shareholder’s funds	2.06
Other shareholders funds	1.5
Financial expenses	1.4
Enterprise value	1.01
Current liabilities	1.01
Total assets	0.77
Sales	0.53
Financial revenue	0.01
Current Liabilities: loans	0.01

➢ **For response variable Company Category**

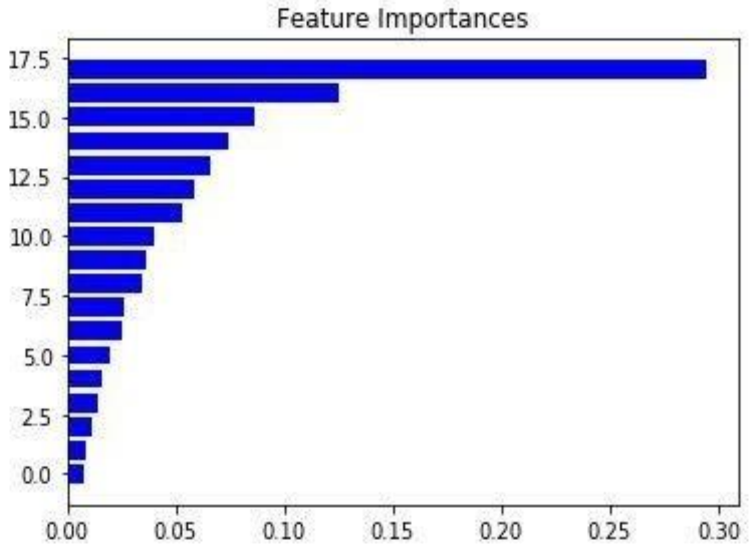


***Figure 3**: Output of Importance function for Company Category*

Most important predictors for response classification of Company Category are as below (in descending order)

•Intangible fixed assets (29.4%)

•Tangible fixed assets (12.4%)

•Current Assets (8.6%)

Table 3: *List of all predictors with their percentage of impact on classification of Company Category:*

Name of the predictor	Impact on the response (%)
Intangible fixed assets	29.48
Tangible fixed assets	12.48
Current assets	8.63
Current assets: stocks	7.45
Other fixed assets	6.59
Current liabilities	5.85
Total assets	5.33
Other current assets	4.05
Shareholders funds	3.67
Current assets: debtors	3.44
Shareholder funds: capital	2.65
Cash & cash equivalent	2.5
Current Liabilities: loans	1.99
Working capital	1.65
Other shareholders funds	1.42
Current Liabilities: creditors	1.19
Total shareh. funds & liab.	0.83
Other current liabilities	0.81

3. **Scaling and Normalization**: - Once the process of variable selection was completed it was time to scale and normalize the data to avoid the problem of bias towards any particular output.

4. **Principal Component Analysis: -** One the predictors were selected we wanted to reduce the number of predictors to avoid overfitting due to more flexibility. Thus, we used Principal Component Analysis (PCA) before fitting our models which helped us represent our model with

less number of predictors. But we still needed to find the minimum number of predictors that would help us represent most of our data without overfitting. Thus, we experimented with different number of predictors to give the best possible accuracy score and the number of components which gave the highest accuracy score was selected.

MODEL SELECTION

We ran multiple models on our dataset to evaluate what fits our data to make accurate prediction. We had two regression and a classification problem statement. We narrowed our research to the following three models: -

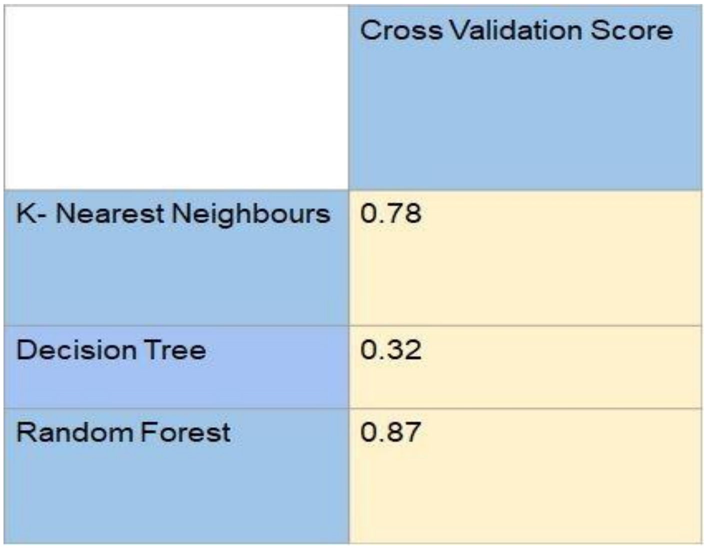
K-Nearest Neighbors: is a completely non-parametric approach. To make predictions for an observation X = x, it assumes K training observations closest to the observation x are identified. Then X is assigned to the class to which these observations belong. Since KNN is a completely non-parametric approach, no assumptions are made about the shape of the decision boundary and therefore this approach dominates when the decision boundary is highly non-linear.

Decision Tree: is a model that predicts the value of a target variable based on several input variables. It uses information gain as a parameter to decide the root notes and split the data. One of the main advantages of using decision tree is the interpretability of the model, every possible scenario from a decision finds representation by a clear fork and node, enabling viewing all possible solutions clearly in a single view.

Random Forest: It is an ensemble method in which a classifier is constructed by combining several different Independent base classifiers. It is a collection of decision trees; the idea is that each of the individual trees in a random forest should do reasonably well at predicting the target values in the training set. Random forests tend to be more accurate than decision trees. There is significantly lesser chance of overfitting in Random forests as it averages several decision trees to create the model

FINDINGS

REGRESSION: OPERATING REVENUE/TURNOVER



	Cross Validation Score
K- Nearest Neighbours	0.78
Decision Tree	0.32
Random Forest	0.87

Figure 4: *Figure on the left shows the cross-validation score for respective models with Operating Revenue/ Turnover as the response variable.*

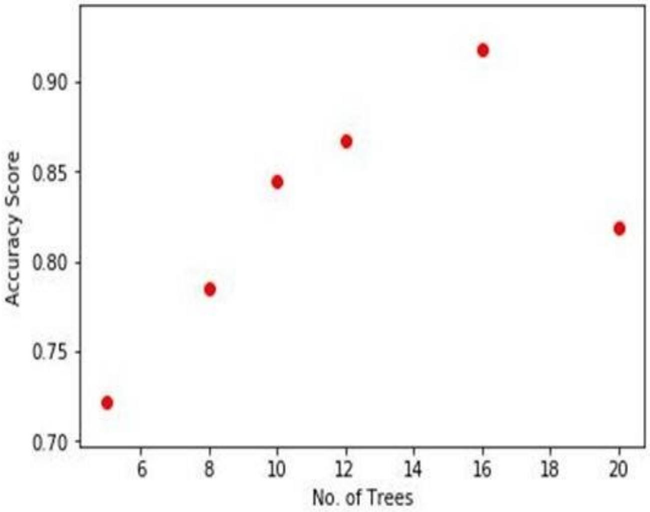
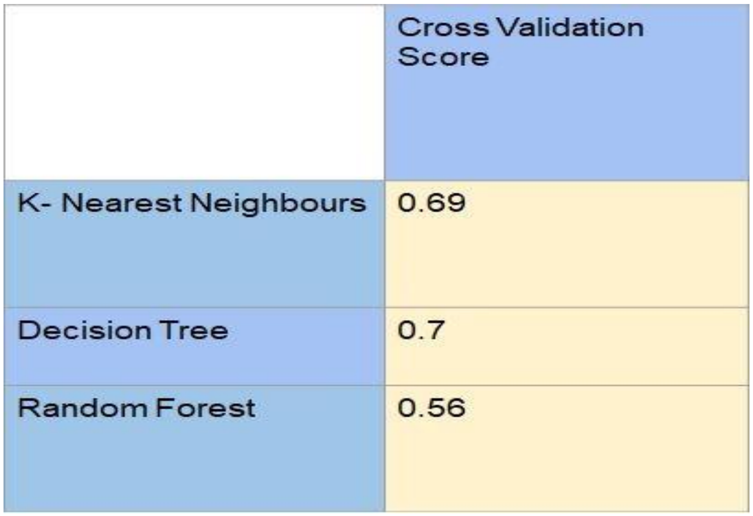


Figure 5: *Figure on the right indicates the predictive accuracy (R-Squared score) by changing the hyper parameters (No of trees in decision tree).*

The cross-validation scores for K-Nearest Neighbors, Decision Tree and Random Forest are given in the above table. By observing the accuracy scores, we concluded that Random Forest is the best optimal solution for our data since it has the highest accuracy of 87 percent. But running Random Forest requires time and resources. Thus, it is the best possible solution in case of no constraints with respect to time and resources.

REGRESSION: GROSS PROFIT



	Cross Validation Score
K- Nearest Neighbours	0.69
Decision Tree	0.7
Random Forest	0.56

Figure 6: *Figure shows the cross-validation score for respective models with Gross Profit as the response variable.*

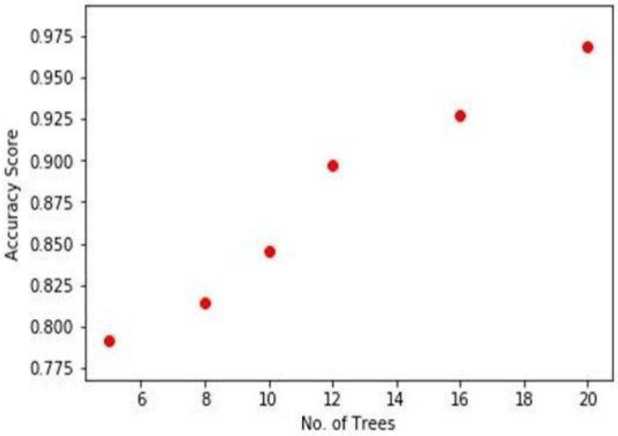
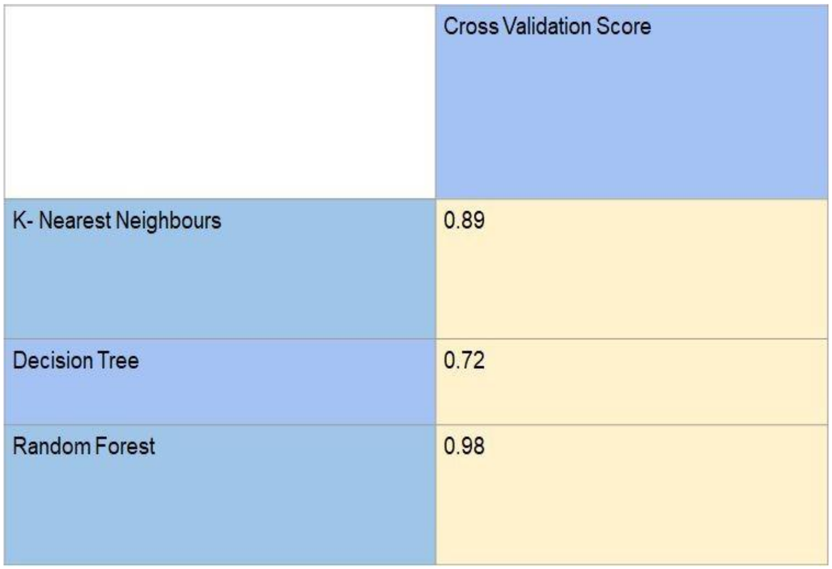


Figure 7: *Figure on the right indicates the predictive accuracy (R-Squared score) by changing the hyper parameters (No of trees in decision tree).*

The cross-validation scores for K-Nearest Neighbors, Decision Tree and Random Forest are given in the above table. By observing the accuracy scores, we concluded that K-Nearest Neighbors is the best optimal solution for our data since it has the highest accuracy of 69%. Also, this leads to a further conclusion that the decision boundary of our data for gross profit maybe nonlinear. Principal Component Analysis (PCA) was used to find the best number of components which lead to the highest accuracy score which was found to be 3 in this scenario. We tested the model for different number of components 1,3,8,13. The one with the highest cross validation score was selected.

CLASSIFICATION: COMPANY CATEGORY



	Cross Validation Score
K- Nearest Neighbours	0.89
Decision Tree	0.72
Random Forest	0.98

Figure 8*: The above shows the cross-validation score for respective models with Company Category as the response variable.*

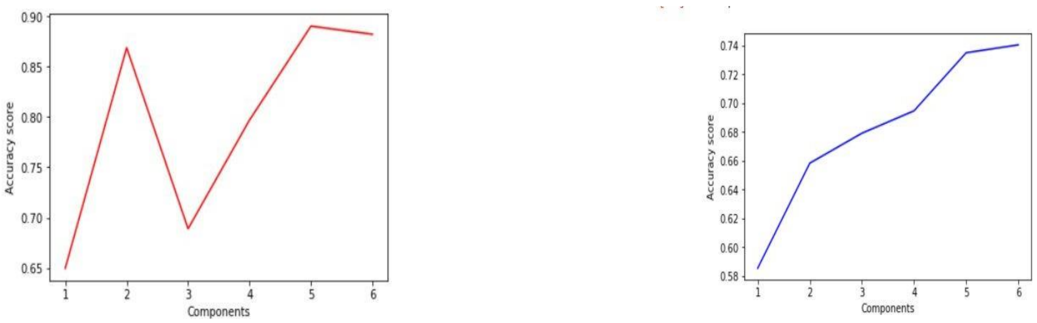


Figure 9*: Figure on the left shows the accuracy score for different number of components for Decision Tree Classifier. Figure on the right shows the accuracy score for different number of components for K-Nearest Neighbor.*

The cross-validation scores for K-Nearest Neighbors, Decision Tree and Random Forest are given in the above table. By observing the accuracy scores, we concluded that Random Forest is the best optimal solution for our data. But accuracy score is too high (98 percent) which also indicates overfitting. With the help of our process of selection of variables we were eliminate the major factors of overfitting. But this high R-squared scores indicate high multi-collinearity between predictors while transforming them using the Principal Component Analysis (PCA).

CHALLENGES

- Dealing with real world data brings a lot more challenges than any other simulated or already preprocessed data. First and foremost being getting the access to data itself from the WRDS.
- Background study of financial terms and the European economic situations was required to better understand the fields and the data.
- Making data compatible with anaconda packages. For this data was transformed into comma separated value format files.
- Running huge computations on local computers was also another big challenge as the algorithms would crash down many a times.

CONCLUSIONS

- **Variables:** Dealing with the real-world data comes with its own pros and cons. It really unfolds the difficulties that are hidden in the simulated data. It was interesting to know and experience the theoretical concepts of data analysis in practical. The first and foremost thing that we need to learn as a data scientist is the data itself, what are the attributes, what do they mean both literally and figuratively. This obviously requires as much background study of the data and its attributes as it requires to understand our problem statement. On getting on to the analysis, we realized that while some predictor variables may not impact the response variables at all, others might not impact it independently but can have significant impact their interaction with other variables is considered. From our data, we visualized the theoretical fact that using all or maximum of our predictor variables might not actually lead to higher accuracy of the models. On the other hand, doing so definitely covers more and more variance in our data and in turn the R^2 scores get higher and higher. It came as an empirical evidence that higher R^2 score does not necessarily imply better accuracy of the model.

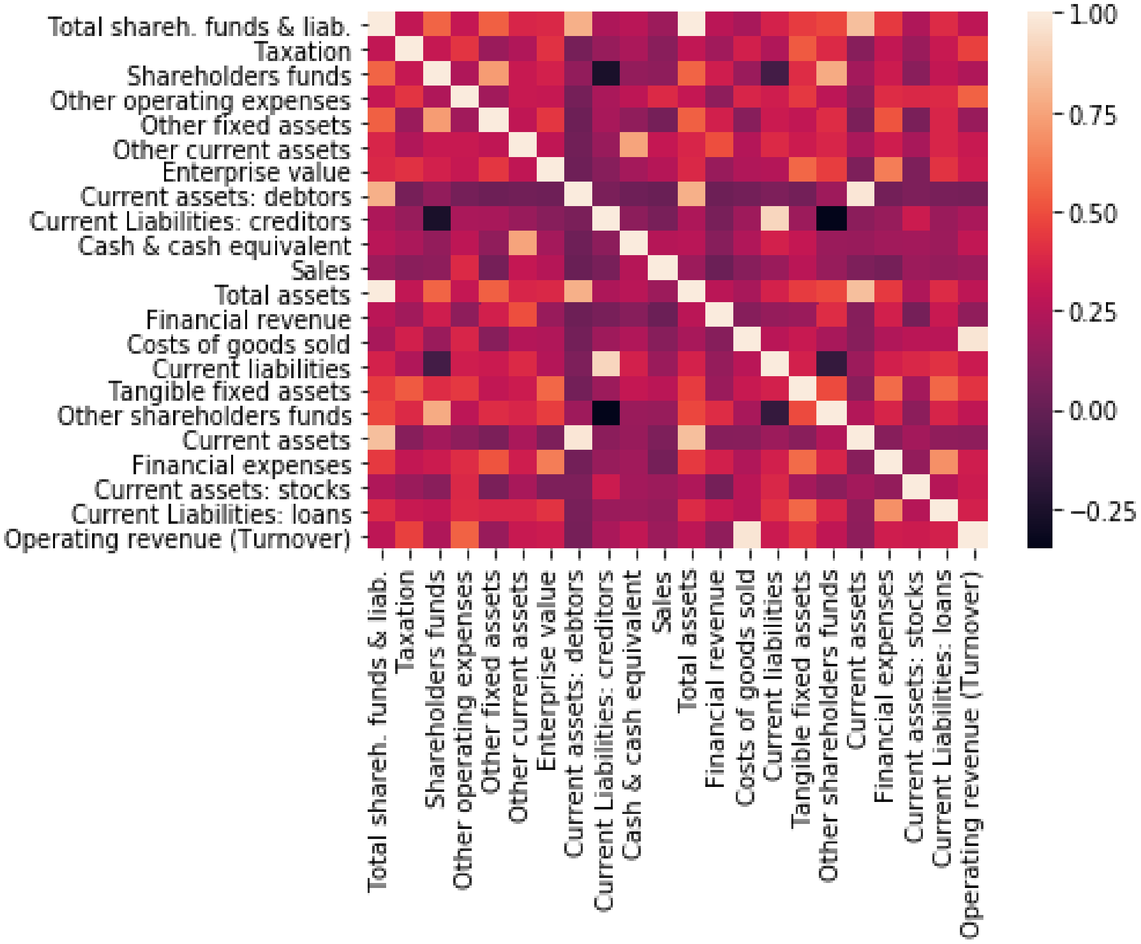


***Figure 10**: Correlation matrix between all the variables in the dataset*

➢ Models:

CLASSIFICATION: The classification problem we had was to classify the firm into categories based on company size. There were four different classes for this namely Small, Medium, Large, Very Large. We built three different models for this classification which were KNN classifiers, Decision Trees and Random Forest. For all the three models the initial approach was to do the model specific preprocessing. It included feature selection, filtering and replacement of missing values. Again, these processes were done iteratively for each model also as to figure out the best suitable combination of test and train split ratio. Once the models were built, they were evaluated based on their cross-validation scores. Very clearly, Random Forest turned out to be the best approach for our classification problem which was clear with the highest cross validation score.

REGRESSION: We had two regression problems, one had “Gross Profit” as the response variable and the other had “Turnover”. For both these problems three different models were built, with each model built iteratively to pick the best combination of test and train split ratio. Then again, the best model was picked based on the cross-validation scores. For the Turnover, random forest turned out to be the best model whereas for the Gross Profit, both decision tree and KNN gave similar results with decision trees being slightly better at the accuracy. This was again evident of the data analysis concept that it is not necessary that a single model give the best results for all response variables and hence we should equally focus on learning picking up between models and not just building models.

➢ Accuracy Metrics:

Confusion Matrix - For the classification of firms into four different categories, i.e small, medium, large and very large, the Random Forest turned out to be the best with an accuracy score of about 92 percent.

Random Forest

Accuracy score - 0.92870739087550361

Confusion Matrix –

```
array ([[11014,   171,    11,    48],
        [  281, 41950,    99,     6],
        [   99,   144, 63154,     1],
        [  340,    66,    11, 2239]], dtype=int64)
```

The next best model for the classification of firms was Decision Tree, it gave an accuracy of around 90 percent.

Decision Tree

Accuracy Score - 0.89467826871959474

Confusion Matrix –

```
            array ([[10655,   266,    55,   268],
                    [  257, 41751,   261,    67],
                    [   47,   223, 63117,    11],
                    [  275,    82,    21, 2278]], dtype=int64)
```

The least good of all three models, but still significant enough, turned out to be the KNN model, with around 80 percent accuracy.

Random Forest

Accuracy score - 0.92870739087550361

Confusion Matrix –

```
  array ([[11014,   171,    11,    48],
                              [  281, 41950,    99,     6],
                              [   99,   144, 63154,     1],
                              [  340,    66,    11, 2239]], dtype=int64)
```

Future Scope

Our project didn’t take account the data for entire Europe due to constraint on time and resources. In the future with the availability of resources we plan to take maximum amount of data for our training set. Also, we split our data in train, test and validation sets but in future we plan to train our data on specific countries and test our model on other countries just to see the accuracy score and the level to which our models overfits on the training data.
In future, we also plan to employ neural networks to have time series analysis of our data over the years in our data and analyses the fluctuations and trends over time. At the same time, we plan to employ deep learning techniques to predict and classify our outcome more accurately and find the most optimal model suited for our dataset.

BIBLIOGRAPHY

Data: https://wrds-web.wharton.upenn.edu/wrds/index.cfm

We have used Mendeley as our citation manager.

Dhingra, S., Ottaviano, G., Sampson, T., & Reenen, J. Van. (n.d.). The consequences of Brexit for UK trade and living standards.

Bartram, S. M., & Wang, Y. (2015). European financial market dependence: An industry analysis, *59*, 146–163. https://doi.org/10.1016/j.jbankfin.2015.06.002

Pellicer, T. M., Pellicer, E., Eaton, D., Pellicer, T. M., Pellicer, E., & Eaton, D. (2014). A macroeconomic regression analysis of the European construction industry. https://doi.org/10.1108/09699980911002584

CODE

https://github.com/omkarpandit24/Statistical-Learning-Assignments-in-R/tree/master/Analysis%20of%20European%20Financial%20Market