DA3-A3

Abigail Chen

Research Question

The goal of this case study is to find out which are the fast growing companies using the **Bisnode firms data**(https://osf.io/3qyut/). We will be building models to predict the fast growing firms. For this case study, we will be focus on the year 2012, with the cross-section of the companies to check whether they are fast growing or not.

Introduction

The main business question this project seeks to answer is whether a firm has been growing fast in the consecutive two years. The classification model for prediction is built using various variables like the company features, balance sheets, HR details and other financial data. The case study focuses on the companies for the years from 2010 to 2015 zooming in on the firms that have high growth rate for two years from 2012 to 2014. to build a prediction model which can support individuals in their investment decisions in choosing between fast and non-fast growing firms.

For this case study, we used 7 different models including OLS, LASSO, Random Forest and OLS logit To classify firms in the mentioned categories, a loss function which quantifies the consequences of the decisions that are driven by the prediction was required (Gabors, 2021). The loss function has two values, one is a loss due to the false negative and a loss due to the false positive. For this purpose we considered these features of the companies and build 7 different models which are OLS, LASSO, Random Forest and OLS Logit. The data comes from the Bisnode, a company that offers decision support in forms of digital business, marketing and credit information.

Summary

We had to use various models, such as random forest, lasso, logit and ols. Here, we can see that the Lasso has the lowest RMSE, at 0.2606437. And the one with the highest AUC is the random forest at 0.7531707 . The random forest has the second lowest RMSE, at 0.2611369.

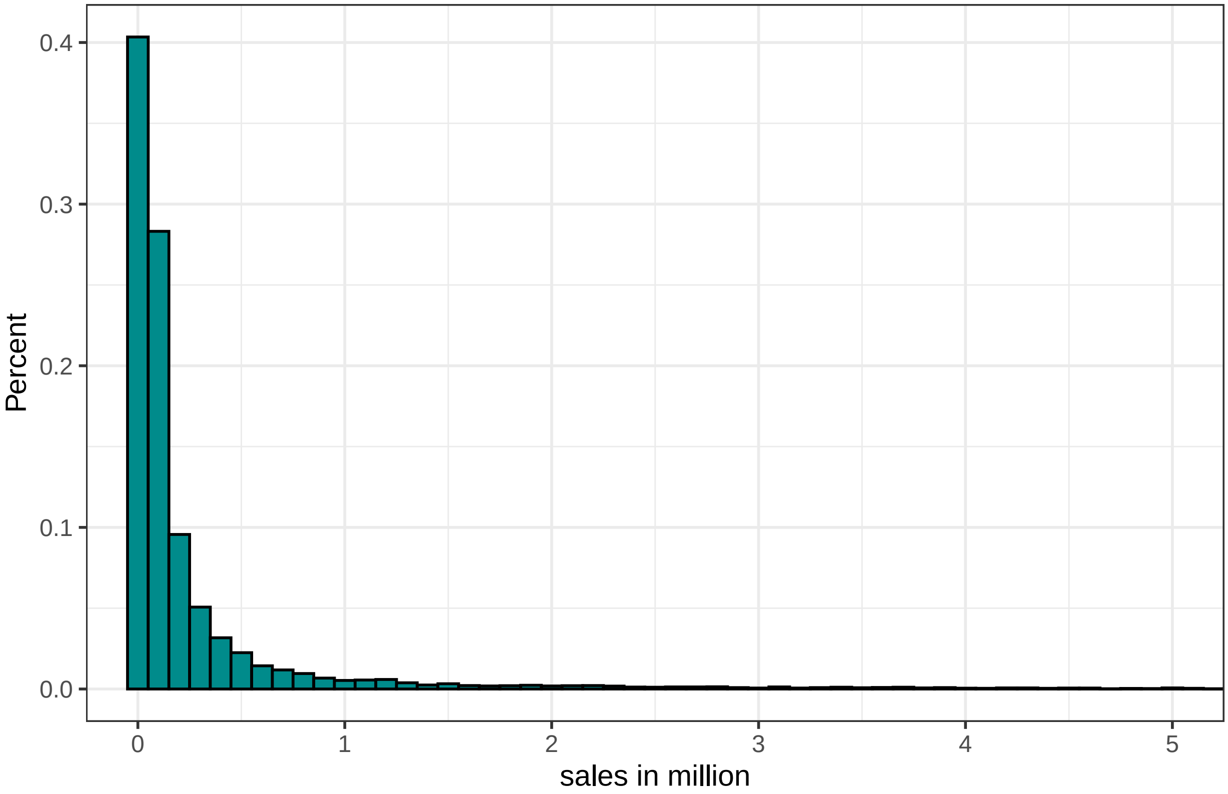
The Dataset

After loading the dataset, we see that there are 287,829 observations with 48 variables. We will then be fixing the variables to the correct formats and work with the missing variables by imputation, dropping, munging and removing null variables.

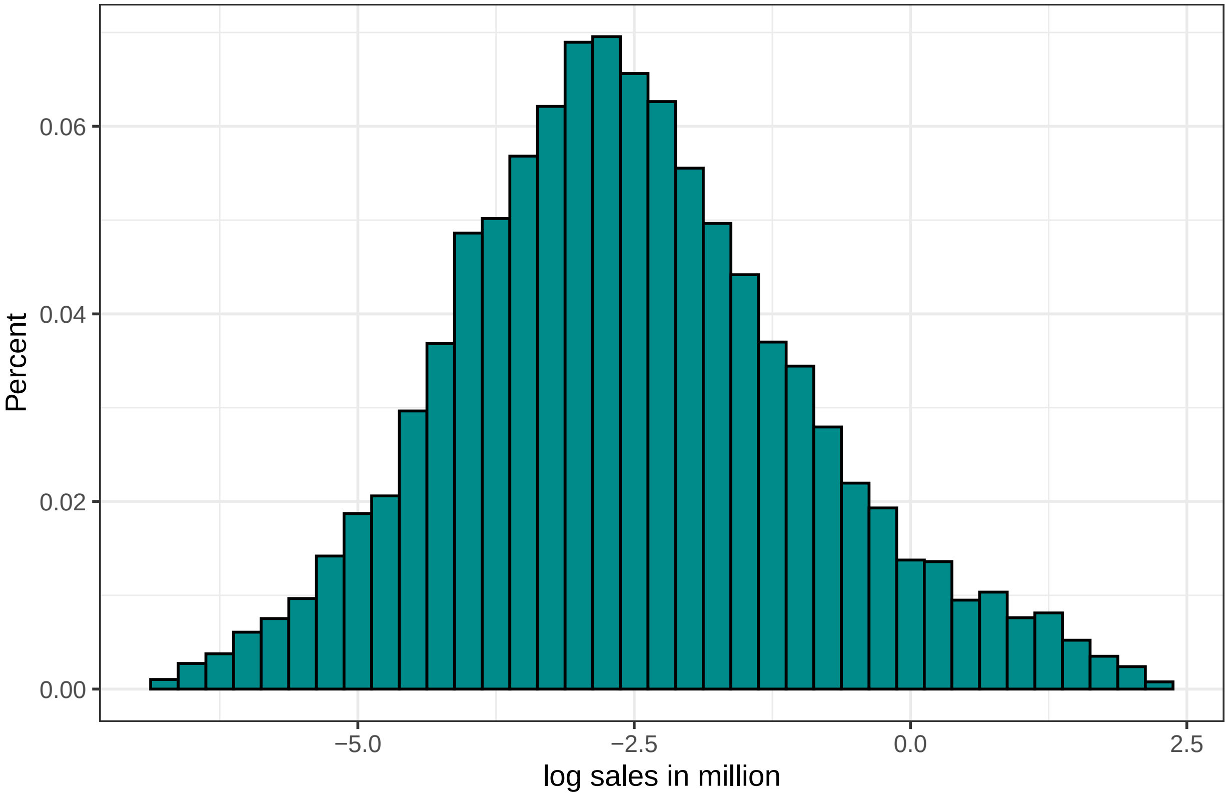
Label Engineering

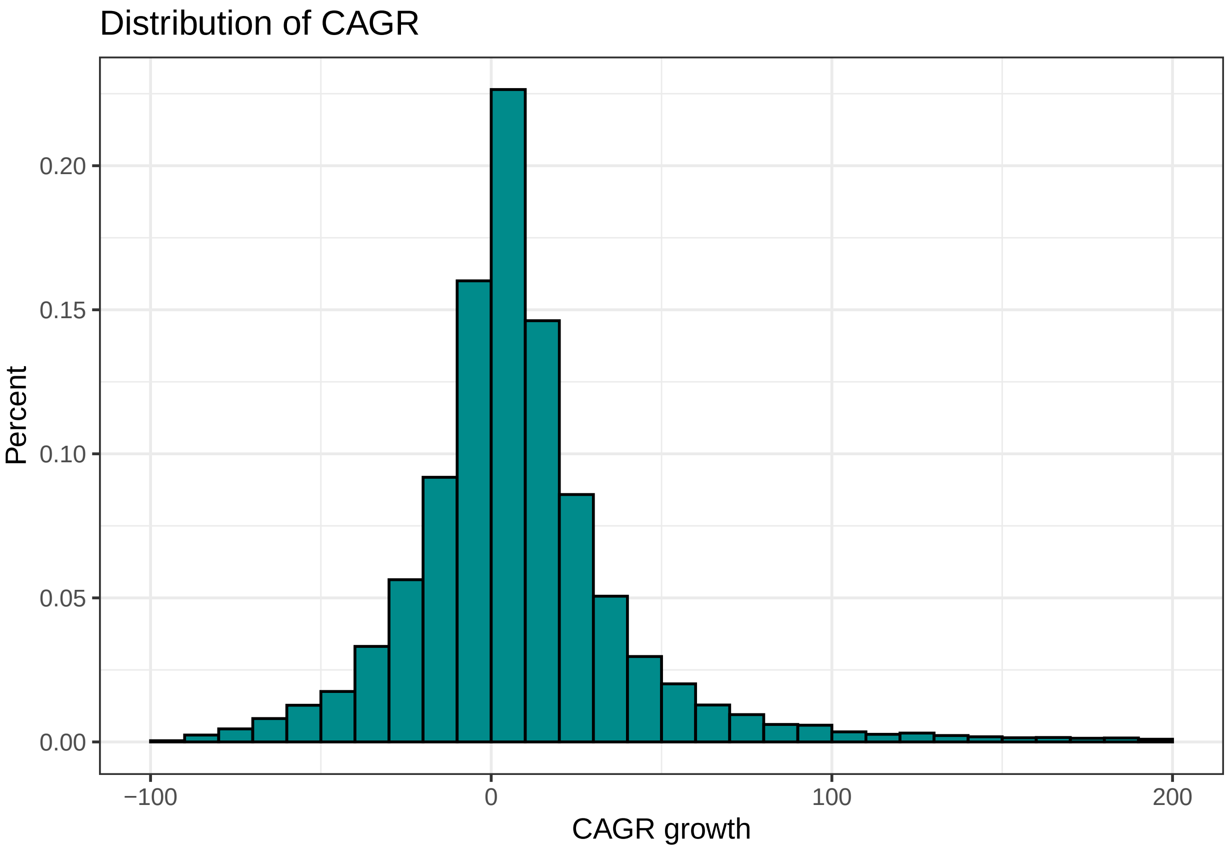
We need to first define the y variable. How do we define a fast growing firm? Here we will be using compound annual growth rate (CAGR).

Distribution of Sales



Distribution of log Sales



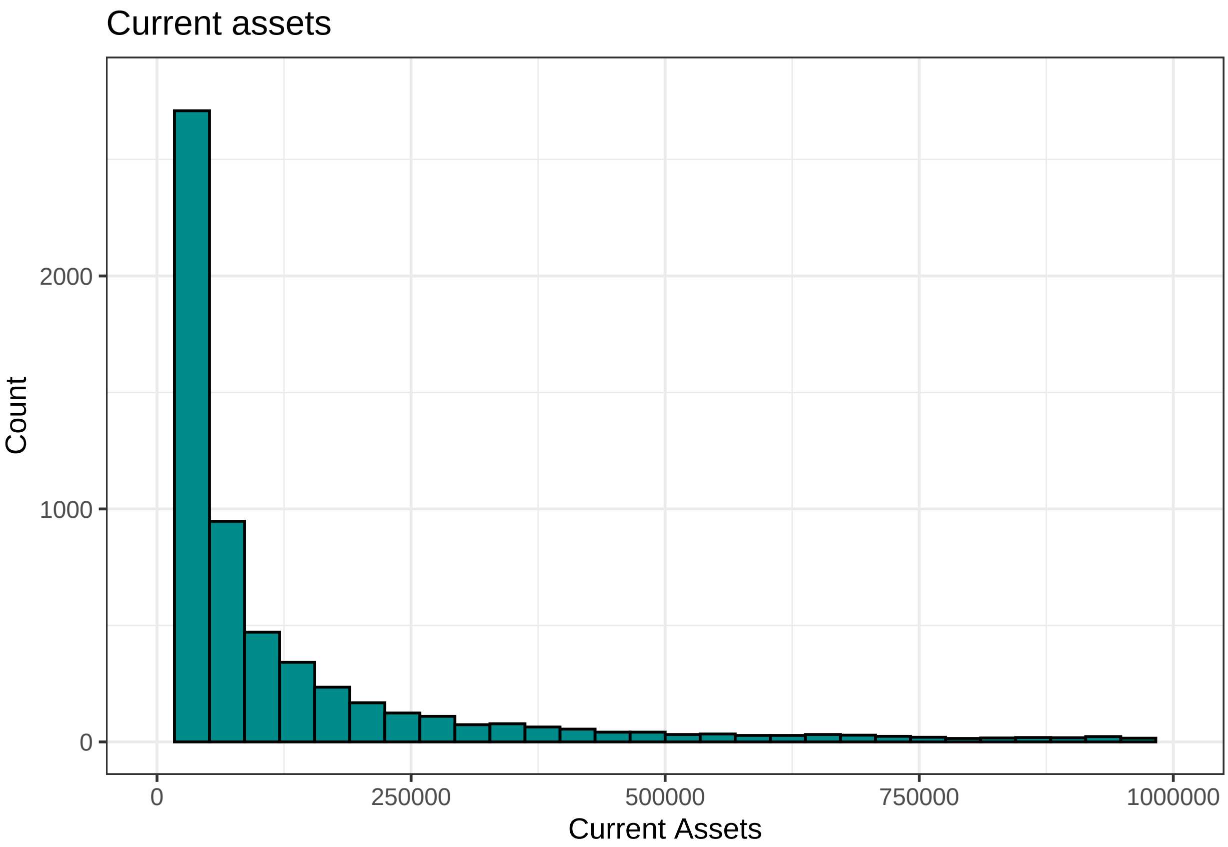


Sample Design

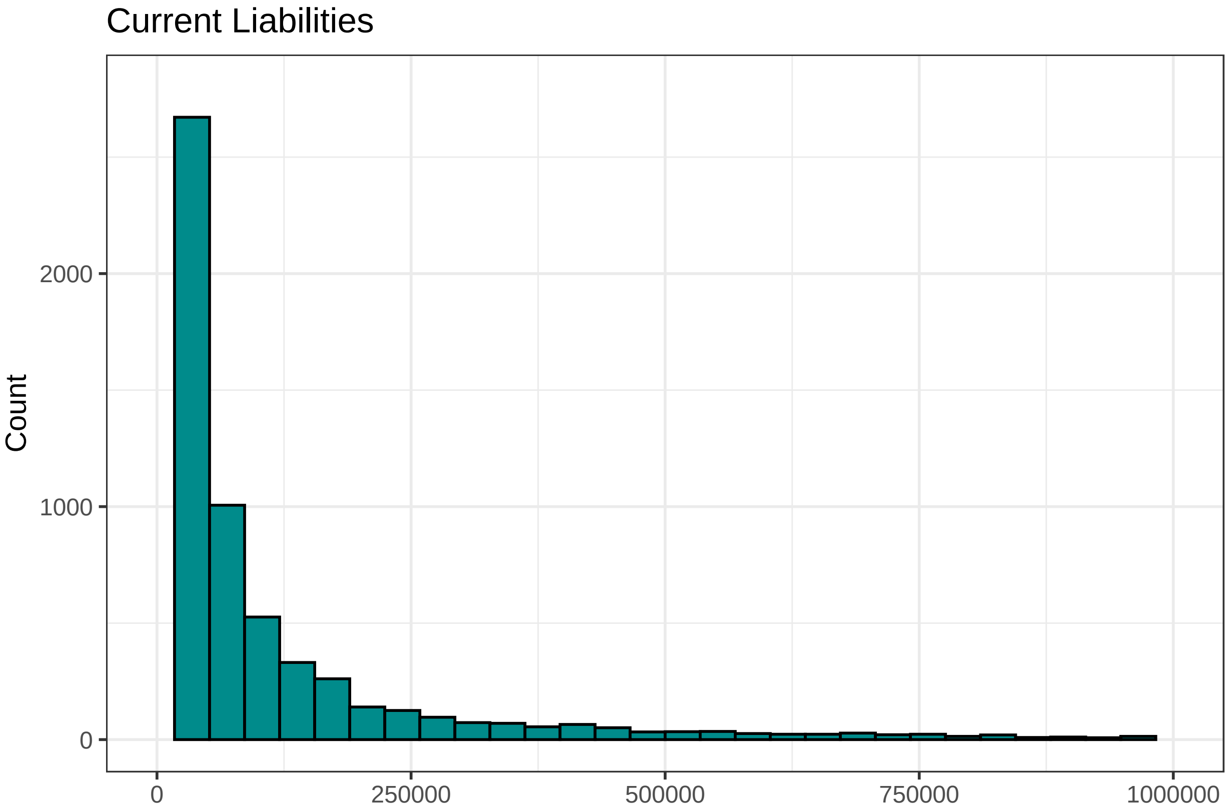
The Bisnode data set contains 287829 observations and 48 variables. This project uses sample data from 2012 to 2014. However, the study was centred on the small and mid-size enterprises captured by 28% of their CAGR sales and companies which had sales between 10 million and 1000 euros in 2012. As a result, sample design concluded with 10462 observations and 117 variables. The main goal of the sample design is to reduce the impact of extreme values. Moreover, the sample design incorporated an alive status filter to ensure that all firms are still operating in the market.

Feature Engineering

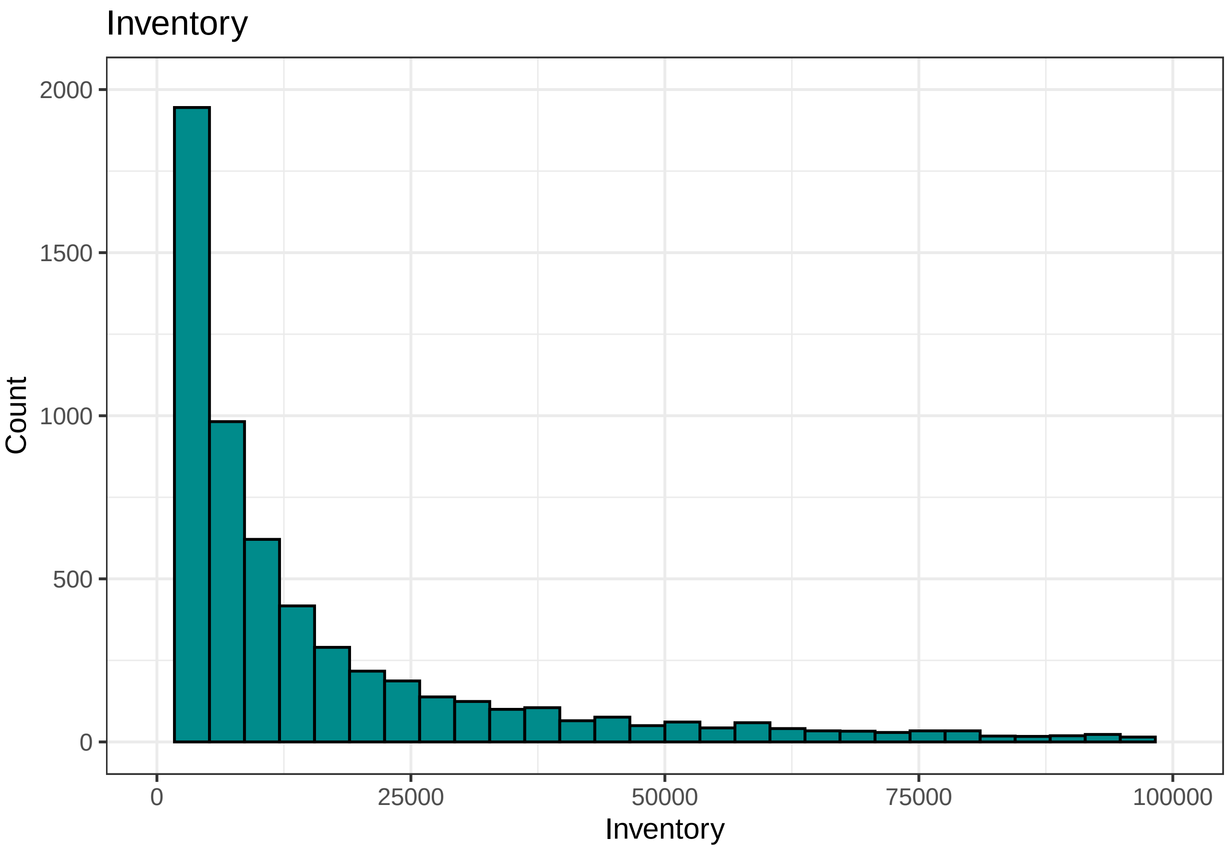
The next task in the case study is feature engineering, which consists of selecting, cleaning and putting the x variables, in proper forms for the model prediction. The variables have different characteristics such as the firm size, financial factors, human resource and others. The main thing about feature engineering deciding what functional forms of variables should be included.

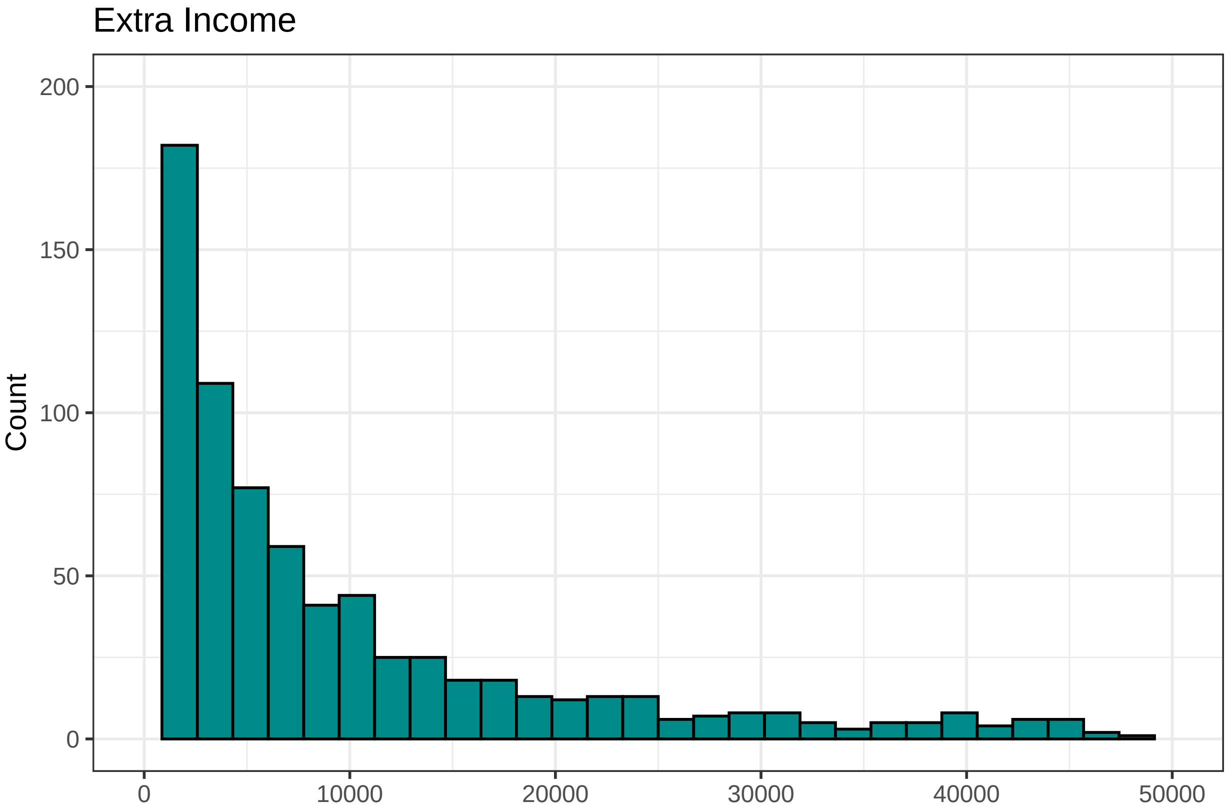


Current assets



Current Liabilities





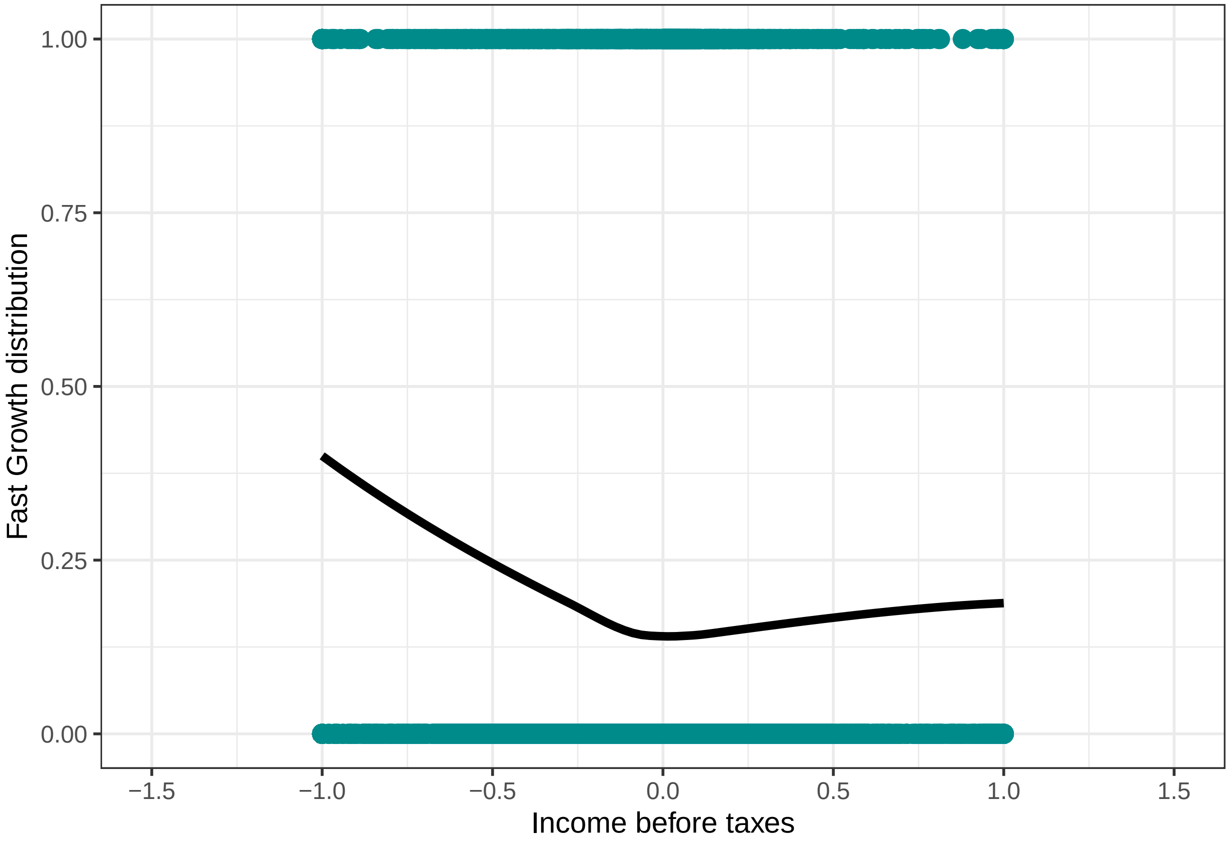
There are various ways to address such values. Based on the the Data Analysis book, we can transform the

Table 1: Firm Growth Summary

Growth	N	Percent
no_fast_growth	9707	92.11
fast_growth	831	7.89

functions to its logarithmic form (ln), group the factor variables or use winsorization. In winsorization, we identify a threshold value for the various variables and substitute the value outside of the threshold with the threshold values itself and finally adding a flag variable. We can also make new variables based on the results of the distributions shown. We can also create new columns for the various profit and loss variables. We can also create a flag variable to select variables which cannot be less than 0.

This figure shows a curve of the relationship between income before tax. for the fast and non fast growing firms.



Variables Description

#Modeling Based on the business research goal, we are to build a prediction model identifying fast and non fast growing companies. For this, we will be using the compound annual growth rate (CAGR) for the two consecutive years 2012-2014. A fast growing company is fast when it does better than the growth rate. Two years was chosen vs one year because it's easier to identify companies as fast vs non fast, compared to just looking at a one year time period. This period allows us to see a better cumulative annual growth rate, where we can see if they can maintain this from the first year to the second year.

Here we can see that the 92% of the companies have no fast growth while the remaining 8% has fast growth.

Set up

The best model will provide the best prediction. To prevent over fitting, the dataset will be divided into two parts in a 80:20 rarion. The 80% is the working dataset and the rest will be the holdout will be 20%. For the training set, we will be using a 5-fold cross validation, by dividing the train data into 5 folds or samples and then choosing based on the average of the 5 CV RMSE result.

Probability Prediction and Model selection

Probability Logit Models

In the case study, we used logit to perform the probability prediction and then choosing the best logit model by crossvalidating and evaluating the model. We used 5 various logit models. The first model is the one with the domain knowledge. Here are the various variables:

In order to do a model comparison we will be using a standardized measure to choose the best model among the various models we will make. We will be using two measures. First is the root mean squared error or RMSE, and the area under the curve or (AUC). The differences in the RMSE we got is quite minimal. Here, we can see that the RMSE is lowest for X3, which is around 0.2618937. And for th AUC, the highest is X3 too, which is at 0.7415733.

	Number.of.predictors	CV.RMSE	CV.AUC
X1	11	0.2650568	0.6841460
X2	18	0.2636046	0.7096458
X3	35	0.2618937	0.7415733
X4	78	0.2619683	0.7397117
X5	152	0.2641646	0.7290488

LASSO MODEL

For this part, we will be including LASSO for logit. Here we can see that LASSO, has the better RMSE. While for AUC, the third model has the best value at 0.7415733.

	Number.of.predictors	CV.RMSE	CV.AUC
X3	35	0.2618937	0.7415733
LASSO	27	0.2606437	0.7178062

Random Forest

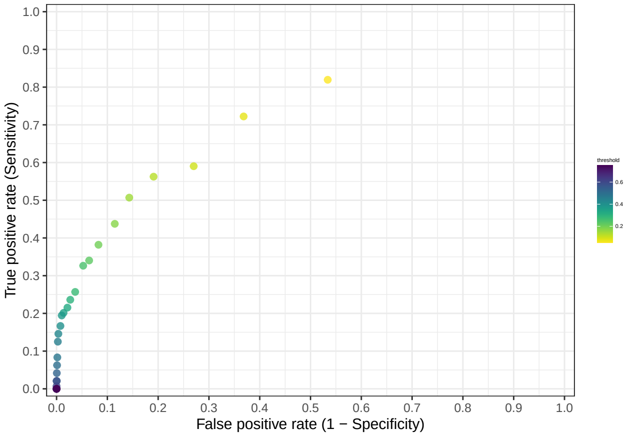
The last model we created is the probabiloty Random Forest, where we will be using the set of variables from the random forest modelling. The variables we will be using are the raw variables, firm variables and HR variables.

Here we can see that the RMSE dont different much, but the random forest gave a lower RMSE at 0.2611369, and a higher AUC at 0.7531707.

	CV.RMSE	CV.AUC
X3	0.2618937	0.7415733
Random_forest	0.2611369	0.7531707

ROC Curve

What is the ROC curve? It can graphically show the trade-off between the false positive and false negative when we use various classification thresholds to the probability predictions. We will then we using this for our selected model.



Finding the Optimal Classification Threshold

Lastly we will look for the loss function. This give value to the consequences of decisions that are driven by prediction models. A loss function can have two values, from the false negative and the false positive. This will help us in decision making. We need to get the threshold value for the classification which will give the smallest loss.

Here, we can see that the Lasso has the lowest RMSE, at 0.2606437. And the one with the highest AUC is the random forest at 0.7531707 . The random forest has the second lowest RMSE, at 0.2611369.

	Number.of.predictors	CV.RMSE	CV.AUC	CV.threshold	CV.expected.Loss
Logit X1	11	0.2650568	0.6841460	0.3375201	0.1519391
Logit X3	35	0.2618937	0.7415733	0.3555382	0.1505159
Logit LASSO	27	0.2606437	0.7178062	Inf	0.1549875
RF probability	36	0.2611369	0.7531707	0.3411077	0.1511090

Best Model based on Expected Loss

Here we can see that the Random Forest has the lowest expected loss and RMSE, compared to the other models. Random forest also has the lowest expected loss.

	no_fast_growth	fast_growth
no_fast_growth	1933	120
fast_growth	30	24