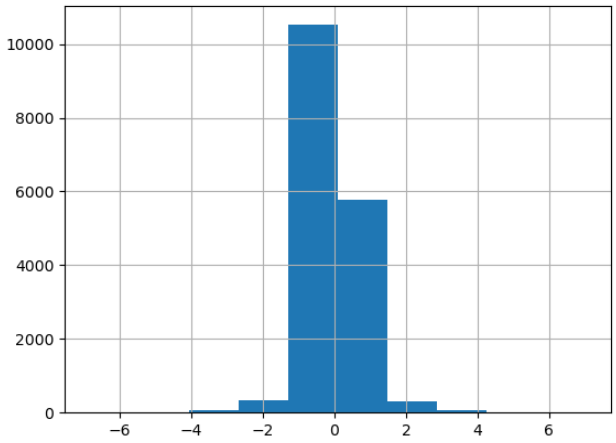
Technical report

In this report, I discuss technical details of how I predicted the fast-growing firms in the market in 2012 based on available firm level characteristics using bisnodes-firms data. Codes and outcomes are available at: https://github.com/Dilnovoz/Data-Analysis-3/tree/main/Assignment_3

Target labelling

Growth of the firm are measured based on log sales changes over 1 year period. Sales growth are replaced with 0 if the firm just started the operation in the market. According to the given histogram, sales growth rate from 2012 to 2013 was negative for more than half of the firms with average growth rate of 4 percent. Further our target variable: firms are assigned as fast growing firm if their sales growth is above 38.7 percent from 2012 to 2013 or if their growth are in top 15 percentile. Other measurements of growth rates like wage growth, exporting products, labor growth are not chosen as these variables are noisily measured in the dataset.



Feature engineering

Features with most missing values (COGS, finished_prod, net_dom_sales, net_exp_sales, wages) are dropped from the analyses. Industry variable had a lot of subgroups of industries that are regrouped into 12 aggregate industries so in analyses their effect is clearer and, in each category, there are enough number of firms to analyze. Different type of asset variables as well information from P&L statement where have negative values are replaced with 0 as assets and expenditures cannot be negative value and flag variables for these changes are also introduced to the model accordingly. Average number of labor variable had also around 15 percent missing observations and missing values are replaced with sample average and to control for possible nonrandomness of the labor missing, flag variable is introduced to the analyses. If age of the CEO is really young,

younger than 25 or old, older than 75, they are replaced with 25 and 75 accordingly as it is nearly impossible to have a CEO of this age. Furthermore, CEOs who are younger than 40 years old are categorized as young CEOs.

Sample restriction

Samples are restricted to the firms operated in 2012 and 2013 and sales volume of between 1000 and 1 mln. If some variables had very few numbers of missing observations, these observations are dropped from the sample. At the end I had almost 17 thousand firms which 14482 of them are not fast-growing firms to prediction analyses.

```
data.High_growth.value_counts()
```

```
0    14482
1     2340
Name: High_growth, dtype: int64
```

Model selection and prediction

For the logit regression estimation, interaction terms of relevant variables as well as quadratic term of firms' age and sales volume are introduced in this model as there is parametric relationship is assumed when we introduce the features to predict the outcome. Thus nonmonotonicity of the relationship between introduced features and target variable are corrected with these additional variables. Further, Logit model with Lasso function assisted to remove irrelevant variables from the model. According to Logit with Lasso model most useful variables to predict the sales growth are, sales amount in 2012, and age of the firm as well as the industries firm operated in. As we know ballsheet_notfullyear indicates that the firm only operated only few months in this year it is negatively associated with sales growth volume while new firms that are existed in the market since the beginning of the year are more likely to be a fast growing firm in 2013.

Table 1

Variable	coefficient	Importance
balsheet_notfullyear	-0.57529	0.575295
sales_mil_log_sq	0.304656	0.304656
sales_mil_log	-0.25089	0.250889
C(ind2_cat)[T.56.0]	-0.18756	0.187564
age	-0.16178	0.161779
ceo_age	-0.13858	0.138579
new	0.135952	0.135952
C(ind2_cat)[T.33.0]	0.119475	0.119475

material_exp_pl	0.118943	0.118943