Mukhamejan Assan
Assignment 3

Finding fast-growing firms

This project aims to predict the fast-growing firms. I defined it to be a company that is in the top 25% by sales growth in both years 2013 and 2014. I later used 8 models for this task: 5 logit models, LASSO, and Random Forest.

The data preparation process involved filtering a dataset to include only observations from 2010 to 2014. Additionally, variables that had more than 90% missing values were dropped. The target variable was engineered based on the percentage change in sales. Firms were classified as "fast-growing" if their percentage change in sales exceeded a certain threshold. To identify fast-growing companies, the 75th percentile of sales growth in both 2012-2013 and 2013-2014 was calculated. Finally, the data was filtered to include only observations where the year was 2012 and the company was alive, and where the sales value was between 1000 euros and 10 million euros. The total number of companies was 21,723 and the number of fast-growing companies was 2,531, representing 11.65% of all companies.

In total, five logistic regression models were created, one for each of the features X1 to X5. The logistic regression models were trained on the training set, and then evaluated using 5-fold cross-validation. The performance of each model was evaluated using two metrics: root mean squared error (RMSE) and area under the receiver operating characteristic (ROC) curve (AUC).

In addition to the five logistic regression models, a LASSO model was used to select the most important variables from the logitvars feature set. The LASSO model was trained on the training set and then used to select the variables with the highest coefficients. These selected variables were then used to train a logistic regression model, which was also evaluated using 5-fold cross-validation.

In this project, we are particularly interested in minimizing the number of false positives and false negatives. A false positive occurs when the model incorrectly predicts a fast-growing firm when in reality it is not growing fast, while a false negative occurs when the model incorrectly predicts a non-fast-growing firm when in reality it is growing fast.
To quantify the cost of each type of error, we need to assign a weight to each type of misclassification. In this project, I use the median sales growth rate of the fast-growing firms and non-fast-growing firms to assign weights to the two types of errors. Specifically, I define the cost of a false positive as 1 + median sales growth rate of fast-growing firms, and the cost of a false negative as 1 + median sales growth rate of non-fast-growing firms.

I then calculate the expected loss of each model, which takes into account the cost of each type of error as well as the prevalence of fast-growing firms in the dataset. By minimizing the expected loss, I aim to strike a balance between the two types of errors, while also taking into account the fact that fast-growing firms are relatively rare in the dataset.

The results show that the random forest classifier performs the best among all the models tested. It has the lowest root mean squared error (RMSE), the highest area under the curve (AUC), and the least expected loss. Additionally, the number of predictors used in the random forest model is reasonable compared to the other models tested.

Model	# Features	RMSE	AUC	Expected loss
X2	18	0.27	0.93	0.16
LASSO	118	0.24	0.95	0.22
Random Forest	45	0.20	0.98	0.08

The confusion matrix for the random forest model shows that it accurately predicted 3277 out of 3361 (97.5%) non-fast-growing companies and 303 out of 447 (67.8%) fast-growing companies. On the other hand, the confusion matrix for Model 2 shows that it predicted 2988 out of 3361 (88.9%) non-fast-growing companies and 406 out of 447 (90.8%) fast-growing companies.

Overall, the results suggest that the random forest model can accurately predict whether a company will be fast-growing or not based on the given predictors, and it outperforms other models tested in this assignment.