Analytical Report - DA 03 Homework 3

Submitted by: Hanna Asipovich, MSc'23

1.Introduction

The purpose of this assignment is to find fast growing firms based on the historical dataset bisnode_firms, which is based on all registered companies in 2005-2016 in three selected industries (auto manufacturing, equipment manufacturing, hotels and restaurants) in a small sized EU country. The data and the codes used in this case study are accessible from GitHub repo (https://github.com/HannaCEU/bisnode-firms).

2. Data Analysis and Transformation

The original dataset has 287,829 observations and 48 variables for the period 2005-2016. We are interested in the most recent data, thus, we create a panel dataset for the time period of 2010-2015. In the process of data exploration, I carried out data overview and considered main variables for understanding what could be our definition of **a fast growing firm**. As a final choice, the fast growth target was defined though more than 100% increase in sales from year to year. Firms from the dataset with a 100% increase in sales revenues from the prior year were categorized as fast growing. I have also filtered out the firms within the lower 0.5 and upper 0.95 percentile, as we are interested in the consistently well performing firms.

For additional consideration and prediction models building I created **new variables on**:

- **Sales in millions (sales_mil)** and transformed into a log **(sales_mil_log)** to better capture the distribution of a financial variable. Finally, I included **sales_mil_log_sq** to deal with the negative results.
- **age**: this variable was calculated by subtracting the year the company was established from 2014 as the year the data comes from and also deducting those firms which are just newly created and hence have "0" years; further, the additional variable with the squared term *age2* was also added.
- **total_assets_bs**: was created by adding the values of three variables: *intang_assets*, *curr_assets*, and *fixed_assets*. Negatives were replaced with "0" and flagged.
- **ceo_age**: the variable was generated by subtracting the director's year of birth from 2014; additionaly, three flags were added: (1) if *ceo_age* is less than 25, (2) if *ceo_age* is higher than 75, (3) if *ceo_age* is missing.
- Imputed the mean for **labor_avg under labor_avg_mod** and flagged it.

PART I: Probability prediction

Later I define the predictor sets based on our observations and industry knowledge and accounting. In total I estimate 7 models: 5 logit models, one LASSO, one Random Forest. Using the cleaned dataset and based on our exercise in the class, I included 5 logit models for prediction.

Model 1 is simple with 4 variables, Model 4 is the base model with all variables, without interactions. Model 5 has 2 interactions and 77 predictors. We look into RMSE and AUC for each model in the summary table to find out the best performing one.

	Number.of.predictors	CV.RMSE	CV.AUC
X1	4	0.4059555	0.5666712
X2	10	0.3908663	0.6996859
X3	23	0.3795558	0.7373983
X4	65	0.3734107	0.7676381
X5	77	0.3728663	0.7699817

Model 5 (X5) has the lowest RMSE and the highest AUC. This is the best performing model.

I also build a **LASSO model** for eliminating insignificant predictors, resulting in 69 meaningful predictors. It is slightly better performing in RMSE in comparison to Model 5.

	Number.of.predictors	CV.RMSE	CV.AUC
X5	77	0.3728663	0.7699817
LASSO	69	0.3726361	0.7394006

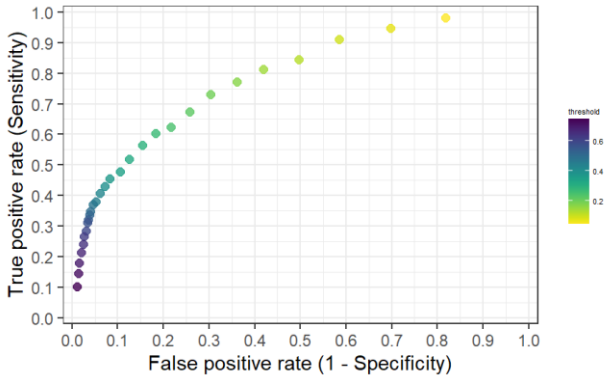
However, **Model 5 is outperforming LASSO in the higher index for the area under the curve (AUC).** It is important for probability of an observation to be well predicted and I will stick to Model 5, as prediction is the core of this exercise.

Based on the comparison table below and earlier choices made, one can say that **Random Forest** shows better results for RMSE and AUC.

	CV.RMSE	CV.AUC
X5	0.3728663	0.7699817
Random_forest	0.3531688	0.8100914

ROC CURVE

One more characteristic for performance of a binary classification model, a Receiver Operating Characteristic (ROC) curve will be reviewed. I used various threshold settings to see relationship between the true positive rate (TPR) and the false positive rate (FPR) and used RF model as best performing. The TPR represents the proportion of positive instances that are correctly classified as positive by the model, while the FPR represents the proportion of negative instances that are incorrectly classified as positive by the model. It is also visualized in the graph.



PART II: Classification

LOSS Function

False positive: The firm which was classified as fast growing but in reality turned out to be making less than 100% growth from earlier. FP=1

False negative: assigning a high performing, 'fast-growing' firm into an underperformer results in lost opportunities. FN=2.

Optimal threshold is about 0.3 for all models. **Model 5** performs best on predicted average losses, by a slight degree better than **Random Forest model.**

```
     Optimal thresholds Expected loss
[1,] 0.2915862          0.4111885
[2,] 0.3013322          0.365751
[3,] 0.3739056          0.324914
[4,] 0.348639           0.3128767
[5,] 0.3302148          0.3119327
[6,] 0.2725363          0.337507
```

PART III: Conclusions

Random Forest is best positioned model to predict probabilities and classify firms into 'yes_fast_growth' and 'no_fast_growth'. Among all 7 models (5 logit models, LASSO, Random Forest), it had the second lowest RMSE (0.35), the highest AUC (0.81), and quite low predicted loss (0.33).