MODELLING FAST-GROWING COMPANIES

Anna-Liisa Altmets, Peeter Niidas, Mahmoud Shoush, Vladislav Samsonov |Group P10|
OUR TEAM

ANNA-LIISA ALTMEES  
Conversion Master in IT

PEETER NIIDAS  
Conversion Master in IT

MAHMOUD SHOUSH  
Ph.D. student in Process Mining

VLADISLAV SAMSONOV  
MSc in Data Science
The project owners are Kristina Lillo and Lennart Kitt.
SEB is a corporate bank founded back in 1856 in Sweden.
THE PROBLEM

To find fast-growing companies using public tax databases. Additionally, if possible, to point out
- stable companies,
- failing companies.

Why?

Banks need to find fast-growing companies ASAP to provide them daily banking solutions and later also provide financing. In other words, that’s a nice way for banking institutions to find new clients.
METHODOLOGY
In the first iteration we decided to label the data according to the changes in turnover as following:

- **growing** companies (turnover change should be positive and mean at least 5% a year);
- **stable** companies (turnover change is between 0 to 5%);
- **failing** companies (turnover change change is negative).
DATA LABELLING | SECOND VERSION

In the second version of labelling we decided to label the data based on the following:

- **yearly turnover growth** has to be at least 20% (eq 5% average quarterly);
- **the continuous average annual employment growth rate (CGR)** has to be in TOP 10% of the companies;
- **the absolute change in the number of employees** has to be more than 10 and the number of employees must at least double compared to the beginning of the period;
- **turnover per employee** has to be in the TOP 10% of the companies.

If a company matches all 4 categories, it is considered as **growing**, none of them as **failing** and all other companies as **stable**.
LABELLING OUTCOME

Labels by turnover:
- Stable: 39.2%
- Growing: 21.8%
- Failing: 5.6%

Labels by 4 parameters:
- Failing: 38.8%
- Stable: 27.7%
- Growing: 0.1%
MODEL BUILDING

SPECIFICATION
Multiclass classification

MODELLING
Selecting best parameters, building the model (XGBoost)

OPTIMISATION
Hyperparameter optimisation using XGBoost and HyperOpt

EVALUATING
Calculating performance metrics like ROC and AUC, assessing the model’s performance
RESULTS
COMBINED I FIRST VERSION

AUC ≅ 0.7530
COMBINED I SECOND VERSION

AUC ≅ 0.8558
SHAP ANALYSIS

- cum_perc_turnover
- turnover_growth
- cum_sum_turnover
- Turnover
- turnover_per_employee
- cum_sum_combined_taxes
- Business_area_by_EMTAK_system
- emp_growth_rate_CGR
- number_of_employees
- Labor_taxes
- All_taxes
- combined_taxes
- ACE

mean(|SHAP value|)

- Failing
- Growing
- Stable
CONCLUSION

● Model trained, growing companies found based on two versions of labelling.

● Performance was better on before the pandemic dataset.

● The multi-labelling method turned out to be more accurate and could learn the actual representation of the data (just as it was expected).
LESSONS LEARNED

REAL DATA
vs synthetic data from lectures

PLANNING
work hours and waiting hours

TEAMWORK
planning and working with new people
THANK YOU!

Any questions? Comments?