H7 Report

Predicting employee salary range based on company details in Estonia.

Kaspar Kipp, Andres Rõõm

Repo link: https://github.com/KasparKipp/ids

Business understanding

Identifying business goals

The Estonian Tax and Customs board provides quarterly overviews of the businesses operating within their jurisdiction including data about their labour taxes and payments (part of spend on workforce, everything that is not income tax and the salary that reaches the employees bank account) and number of employees working in the company at that time. Now everyone can consult that data and as a matter of fact, people have made a business out of making that data available in a convenient matter (inforegister, teatmik).

For the working class, this overview in itself is not really informative without specific domain knowledge in assessing if a potential employer will be able to pay a decent salary in the future and what could the salary range for the people employed in the company be.

Our primary goal is to support new hires in gaining a clearer understanding of expected salaries and future prospects across various companies in Estonia. To achieve this, we are leveraging publicly available data from the Estonian Tax and Customs Board, which provides information such as quarterly taxes paid, turnover, number of employees, and additional relevant attributes.

This data is further enriched with open data from the Estonian E-Business Register to create a more comprehensive dataset. By testing different regression models, we aim to provide predictive insights into future workforce expenditures.

For students, career changers, or individuals transitioning between companies, navigating salary expectations can be challenging. Our solution leverages historical data, helping users make informed decisions about their career paths and salary negotiations.

The proposed machine learning based workforce expenditure prediction models output should be translated into terms of company average salary and give a

meaningful prediction about the future average salary in a company. A reasonable human readable error range (for example Mean-squared error of the model) should be provided alongside the result, which should be small enough for the result to not be rendered meaningless.

Assessing situation

Two data science students completing their project with two laptops, open data from Estonian Tax and Customs and E-Business register, bag of chips and at least 30 hours of expandable time on their hands. The project is anticipated to be completed no later than December 13.
Three main goals of the project are:

- Preprocess and merging public datasets
- Perform exploratory data analysis and feature engineering on historical data
- Test different regression models to predict spend on workforce and to produce a notebook app that given a company name, gathers data from the the web and based on historical data, predicts current salary range for employees in that company in the future.

We consider the project to be a success if we manage to complete at least the first 2 of the steps, the third step might be delayed or canceled due to step 1 and step 2 completion taking longer than expected because of our lack of prior experience in data science projects.

Limitation in calculating expected salaries based on reported workforce taxes is the lack of information regarding how many employees have opted out of the funded pension. For example, in a company with a 3,000€ workforce expense, an employee who contributes the default 2% funded pension would take home 1,729.15€ after all deductions, while an employee who has opted out of the funded pension would take home 1,765.02€. Starting in 2025, Estonians will have the option to increase their funded pension contribution up to 6%. However, due to the absence of data on how many individuals have opted out of the funded pension or might choose to increase their contributions in the coming year, we will base our calculations on the default 2% funded pension for simplicity.

The project will be easy on the wallet, as students are volunteering their time and effort, working on laptops pre-charged during a café visit, and snacking on a bag of chips found at home.

Project specific terminology:

Employer - ‘Company’, ‘Non-profit association’ or ‘Government or state authority’ business entity that employs common workforce and has no limitations to public tax data.
Labour taxes and Payments - Quarterly reported sum statistic by the Estonian Tax and Customs board that includes the total withheld income tax, social tax, contributions to funded

pensions and unemployment insurance premiums for the companies workforce. For example for a single employee with a 3000€ monthly salary, the withheld income tax would be 578.4 €, social tax 990 €, contributions to funded pension 60 €, unemployment insurance premiums would be 24€ by employer and 48€ by employee, all totaling to 1640.4 € and for a quarter it would be 4921.2 €.
Turnover - Total revenue generated during a period, including purchases subject to reverse charge.
VAT - Value added tax.
EMTAK field - Field of activity according to the estonian Classification of Economic Activities in Estonia. Each field has a numeric value. A business entity MUST have a primary EMTAK field dependant in which field of operations a majority of revenue was generated. A business may specify additional EMTAK fields.

Data mining goals

Goal is to combine the data from the Estonian Tax and Customs website with Estonian E-Business register data and find the attributes which impact the company's spend on workforce(expected salary) the most and then make future predictions on historical data. We aim to model expected salaries for various companies, aligning closely with statistical averages while accounting for slight variations. Additionally, our approach seeks to provide deeper insights into potential future salary expectations within the same company.

Data understanding

Data requirements and availability

- Quarterly 'Taxes paid, turnover and size of workforce' reports from Q1 2020 - Q3 2024 (downloaded from https://www.emta.ee/en/business-client/board-news-and-contact/news-press-information-statistics/statistics-and-open-data)
- General data of business entities in Estonia (downloaded from https://avaandmed.ariregister.rik.ee/en/downloading-open-data)

Selection criteria and data descriptions

From the Estonian Tax and Customs Board quarterly reports:

- For legal entity types of 'Company', 'Non-profit association' and 'Government or state authority', these business entity types constitute what we define as Employers for the sake of this project and have general data available without limitations. Quarterly information including:
    - Registry code - unique business entity registry code, used for associating with general data

    - Type - Company type, categorical. We are only interested in ‘Company’, ‘Non-profit association’ and ‘Government or state authority’,
    - Registered in national VAT register - If the business entity was registered as a VAT payer during the quarter
    - Number of employees - Number of employee quarter work hours during the reported quarter, averaged up. Basis for reverse-calculating average salary in company given Labour taxes and Payments
    - Labour taxes and payments - Total labour taxes and payments for the given quarter for all employees
    - Turnover - Turnover for the given quarter
    - EMTAK field
- General data of business entities in Estonia:
    - Company registration date
    - Company name changes
    - Secondary EMTAK fields
    - Company capital - ‘no initial capital deposit’, ‘mandatory capital deposit or < 10k €’, ‘10k € < 100k€’, ‘100k€ <’
    - Company tax debt
    - Company status - In registry, in bankruptcy proceedings etc.

The hierarchical quarterly reports data will be gathered and additional metrics, YoY (Year-over-Year) revenue growth calculated. For experimentation purposes, similar quarterly revenue growth as well for different time frames.
This should give us enough actionable data to train models on.

Problems that we are going to face will be filtering the data since some business entities may cease to exist or stop their business during the periods. This is the sort of pattern we’d like our models to be trained on and finding suitable models and data modeling periods will be the hardest challenge.
Similarly, it might be that companies don’t submit their info to the Tax board on time and the latest quarterly reports are missing a disproportionately large amount of Labour taxes and Payments data.

Verifying data quality

Since our data is hugely dependent on averaging Labour taxes and Payments, the data is skewed because many companies exist that bring in low revenue and have less than 3

employees. This can be overcome by exploring oversampling techniques

	Number of employees
count	58398.000000
mean	8.755060
std	51.441966
min	1.000000
25%	1.000000
50%	2.000000
75%	5.000000
max	4527.000000

The data regarding facts about business entities is luckily all there.

For the Labour Taxes and Payments, the data is there, but seeing that the number of employees quarterly working hours being reported, it is obvious that we are accounting for less than 10% of reported workforce, (as reported by Statistics Estonia https://www.stat.ee/et/uudised/moodunud-aasta-arvestuses-suurenes-tooealiste-inimeste-arv-eestis-25-600-vorra). This might be because this is a self reported statistic, people operating as 'Self-employed person' can't be included as their taxes data is not made public, people work for foreign companies where self reported data can't be verified, people work unofficially and state doesn't have an overview of their salaries and doesn't get taxes from them being employed. This is a limitation we have to deal with.

For all intents and purposes, the aforementioned limitations are not necessarily limitations for our business goals, as feedback about failure of a company to report to the Estonian Tax and Customs board with actionable info (eg. newly founded, no tax history, not following estonian law) why we fail to predict company salary ranges might carry more meaningful information for people prospecting new jobs than the salary range prediction itself, as average salary prediction in company does not necessarily mean that a employer isn't highly qualified and would ask for a bigger salary or vice versa.

Planning your project

Project plan

1. Additional data understanding and preparation. Explore, clean, and preprocess historical data for analysis and modeling. (Kaspar & Andres, 2+2 hours)
2. Pipeline for data preprocessors (Kaspar & Andres, 2+2 hours)
3. Oversampling techniques for model development (Andres 3h)
4. Business specific logic for final interface, helper methods for results interpretations (Kaspar 3h)

5. Model Development and evaluation. Machine learning models to predict workforce spend, including hyperparameter tuning and evaluation metrics. Comparisons between models and different data preprocessors. Validating models on historical data. (Kaspar & Andres, 10 + 10 hours, in person)
6. Interface Design and Development. Explore exporting models or creating a simple UI in Jupyter Notebook (Andres 6h. Kaspar 3h)
7. Model integration and deployment of the final product. Making the results available for users. CI/CD (Kaspar 3h)
8. Testing and feedback gathering. Making the product usable for our peers (Kaspar & Andres, 4 + 4 hours)

Methods and Tools

- Exploratory data analysis and feature engineering - Python, Pandas, NumPy, Seaborn, Jupyter, SMOTE, ADASYN.
- Model development - Python, Pandas, Numpy, Scikit-learn, Seaborn, TensorFlow
- Interface Design - providing MSE, reverse-calculating salary ranges from tax spend, Python, Jupyter, Seaborn, javascript, tbd.

The steps needed to complete the project are outlined in the project plan and need to be completed in an orderly fashion with the exception of testing and feedback gathering, which can start after preliminary model training and should span the rest of the project.

For exploring deep learning/neural networks, additional time might be required to set up a cloud environment, as we don't have access to machines with performant GPU-s.

Iteration of models and figuring out how to overcome some limitations of the data and nature of seasonal fluctuations in employment in some fields might require us to reduce the hierarchical data or provide additional models dependent on prediction period for most accurate results.

Project and project poster have to be finished by 9th December due to team members other obligations.