H4 Job market analysis

A:Karl Joosep Põldsepp

B:Joosep Luts

1.Background

When the time arrived to choose a topic for the project in the “Intruduction to Data Science” course, we arrived at the conclusion, that in this time of many crossroads in our young lives, it matters to choose a path most relevant to ourselves, be it pursuit of career or academic knoweledge. Thus we reached to the job vacancy websites, to extract information about situation in the job market. Which specialications are most in need on nowadays market, how usefull is a university degree and what are the expected wages.

Business Goals

Primary Goal: To analyze LinkedIn job postings to uncover key trends and insights in the job market that can aid university students and job seekers.

Secondary Goals

To understand what skills and qualifications are most in demand in various industries.

To identify patterns in salary based on factors like industry, job title, and location.

Success Criteria

Generation of actionable insights that can be utilized to improve our study, job search and application strategies.

2.Inventory of Resources

Data: The Kaggle LinkedIn dataset, Töötukassa API

Technology: Data analysis and machine learning tools (e.g., Python, Jupyter Notebook, Scikit-Learn).

Assumptions: The data is representative of the current job market.

Constraints: Limited to the data available in the dataset, may not cover all job sectors.

Risks and Contingencies

Risk of Biased Data: The dataset might not represent the global job market accurately.

Contingency Plan: Augment dataset with additional sources, if needed.

Terminology

Common industry-specific terms should be defined and standardized for clarity.

Benefits

Benefits: Enhanced understanding of job market trends, aiding in efficient job search and study paths.

3.Data-Mining Goals

Trend Analysis: Analyze historical trends in job postings, such as fluctuations in job availability across different sectors, changes in required qualifications over time, and variations in job types and locations.

Skill Demand Analysis: Identify key skills and qualifications that are in high demand across various industries. This can help in understanding which skills job seekers should develop.

Data-Mining Success Criteria

Accurate prediction models with high predictive power.

Discovery of significant insights that align with the business goals.

Gathering Data

Type of Data: The project requires data from LinkedIn job postings and Töötukassa API. This includes details like job titles, company information, job descriptions, locations, posted dates, and salary information.

Time Frame: Kaggle Dataset was made by web scraping job postings twice with months in between in 2023. Töötukassa dataframe was made by pulling Töötukassa API multiple times throughout November and start of December. Duplicates were dropped.

Completeness: According to Linkedin dataset author, the database is nearly comprehensive.

Verify Data Availability

A dataset from Kaggle titled "LinkedIn Job Postings" has been identified. It appears to contain the necessary fields and information required for the analysis.

Define Selection Criteria

Data fields important to us, such as job titles, industry, descriptions, and company details, wages, skills.

Quality: Data with minimal missing values or anomalies, dataset must be cleaned, empty fields either filled or predicted using algorithms.

Describing Data

Data Fields: In addition to the basic fields like Job Title, Company Name, Industry, and Location, the dataset may also contain fields such as Employment Type (full-time, part-time, etc.), Experience Level (entry-level, mid-level, etc.), Education Requirements, and Skills Required. These fields can provide deeper insights into job market dynamics.

Size of Dataset: 47 Mb of Linkedin dataset contains a nearly comprehensive record of 33,000+ job postings.

3Mb of Töötukassa ocntaining 3,000 job postings.

Exploring Data

Skill Demand Analysis: Analyze the frequency and patterns of required skills and qualifications mentioned in job descriptions. This can provide insights into the most sought-after competencies in the job market.

Employment Type and Experience Level Trends: Investigate trends related to different types of employment (full-time, part-time, contract) and required experience levels. This can help understand the distribution of job opportunities across different categories.

Verifying Data Quality

Data Source Reliability: Given that the dataset is from a reputable source like Kaggle, and presumably aggregated from LinkedIn postings, it can be considered reliable. Eesti Töötukassa is a quasi-governmental organisation and its database is reliable and updated daily.

Time Frame Consistency: 30+30 hours work from both team members, 10 hours a week November-December.

Project Plan

Data Collection and Cleaning (12 hours total):

Team Member A: 6 hours

Team Member B: 6 hours

Exploratory Data Analysis (14 hours total):

Team Member A: 10 hours

Team Member B: 7 hours

Feature Engineering and Preprocessing (10 hours total):

Team Member A: 5 hours

Team Member B: 5 hours

Model Development and Training (12 hours total):

Team Member A: 3 hours

Team Member B: 9 hours

Evaluation and Refinement (6 hours total):

Team Member A: 3 hours

Team Member B: 3 hours

Poster and Presentation (6 hours total):

Team Member A: 3 hours

Team Member B: 3 hours

Methods and Tools

Data Collection and Cleaning:

Tools: Python, Pandas, Numpy

Focus on addressing missing data and standardizing formats.

Exploratory Data Analysis:

Tools: Python (Matplotlib, Seaborn)

Conduct thorough analysis to understand data distribution and identify patterns.

Feature Engineering and Preprocessing:

Tools: Python (Scikit-learn)

Create new features and transform data for model compatibility.

Model Development and Training:

Tools: Python (Scikit-learn, CatBoost, Random forest)

Experiment with different algorithms, including regression and classification models.

Evaluation and Refinement:

Tools: Python (Scikit-learn)

Use metrics like RMSE, accuracy, precision, recall to evaluate models.

Poster Writing and Presentation:

Tools: PowerPoint

Summarize findings and insights in a clear and accessible format.