E1: WHYD/MKSS - When and How You Die/Millal ja Kuidas Sa Sured

Peeter Pissarenko, Martin Toomiste

Repository: https://github.com/PeeterPissarenko/MKSS

Business Understanding

- **Identifying your business goals**
    - **Background**
        - Life expectancy around the world is increasing but causes of death also vary more.
    - **Business goals**
        - Predict the cause and date of death based on nationality, birth date and sex.
    - **Business success criteria**
        - We want to achieve accuracy above 80% on test data.
- **Assessing your situation**
    - **Inventory of resources**
        - Team: Martin Toomiste and Peeter Pissarenko
        - Data: Several csv files of different views on life expectancy
        - Tools: Python, machine learning libraries (e.g. scikit-learn, pandas, numpy) and visualization libraries (e.g. seaborn)
    - **Requirements, assumptions, and constraints**
        - **Requirements**
            - Sufficiently cleaned data on: life expectancy based on nationality, life expectancy based on sex, causes of death based on nationality, causes of death based on sex, causes of death based on date of birth.
        - **Assumptions**
            - We assume that the data we start with is valid.
        - **Constraints**
            - We have a limited amount of data.
    - **Risks and contingencies**
        - Data is fabricated, invalid and hard to process.
    - **Terminology**
        - There is currently no new terminology used.
    - **Costs and benefits**
        - **Costs**
            - Team members’ time
        - **Benefits**
            - Getting new experience and possibly achieving good models.
- **Defining your data-mining goals**
    - **Data-mining goals**

        - Develop a machine-learning model which predicts a person's death year and cause based on a person's nationality, sex and date of birth.
    - **Data-mining success criteria**
        - Model accuracy should be at least 80%.
        - Model predictions should make sense.

Data Understanding

- **Gathering data**
    - **Outline data requirements**
        - We need data tables that include a person's nationality and life expectancy and at atleast one of the following: date of birth, sex, cause of death. It would be best if the data type is csv. The table should have data from as wide a range of birthdates as possible, but at least 50 years.
    - **Verify data availability**
        - We have found many different tables of data that meet our expectations. Some even more so.
    - **Define selection criteria**
        - From the plethora of datasets present in kaggle, we chose *”Life expectancy around the world👩‍🦳 👴”* as it seems to suit our cause the best. From there we filtered out unneeded datatables from the 41 tables provided. We chose only the most straightforward and informational sets.
- **Describing data**
    - We have the following tables:
        - Annual number of deaths by region.
        - Female to male life expectancy for individuals at ages 0, 15, 45.
        - Deaths in certain age groups per year per nation. 0-4, 5-9,...,95-99,100+.
        - Death rate by cause by nationality.
        - Difference of female to male life expectancy at ages 0, 10, 15, 25, 45 65, 80.
        - Life expectancy by age groups 0, 10, 25, 45, 65, 80.
        - Life expectancy at birth for males and females.
        - Modal age at death for females and males per country.
        - Probability of dying that year among males and females of a certain age group 0, 10, 15, 25, 45, 65, 80.
        - Remaining life expectancy for males and females of a certain age group: 80, 65, 45, 25, 15, 10, 0.
    - Total number of usable columns:3+3+21+9+7+2+2+14+14 = 75.
    - The number of rows can be up to 60000 deferring based on the table selected.
    - Such data is of high value to our cause.
- **Exploring data**

    - all of the tables used have the csv format.
    - different values are separated by commas and null values are just empty like so ,,.
    - Most of the tables contain nationality, country code and a selection of years, so the tables are very compatible.
    - Some values that are predicted are negative, this could lead to problems.
    - Different tables have different year spans. this will lead to more null values later on.
- **Verifying data quality**
    - There are only 9 different death causes in our selection of data, this could lead to too vague death causes in our predictions.
    - As the length of tables differ, the amount of null values will be nontrivial.
    - The overall data is trustworthy as it is acquired from Kaggle and it's source can be traced.

Project Plan

- Plan:
    - Data cleaning
        - 2-3h
        - Assignee: Peeter
        - Replace unnecessary values with Nan
    - Merging files
        - 2-3h
        - Assignee: Martin
        - Merge files into one data set and split them into test and training data
    - Training models
        - 3-4h
        - Assignee: Peeter
        - Train models to predict death year and cause
    - Adding prediction complexity with statistics
        - 3-4h
        - Assignee: Martin
        - Manually editing the output of the models to make sure they make sense
    - Analyzing the results
        - 2-3h
        - Assignee: Martin, Peeter
- Tools:
    - Pandas
    - Numpy
    - scikit-learn models and other helper functions
    - Visualization tools like seaborn