Assignment 02

Task 01: Classification

In this exercise, you have to work with a data of past and current employees of a company ('HR data.csv').Your task is to build a classification model to predict which employee will leave the company and which will continue working (the target column is 'left', where 1 indicates that the person left the company). You can find more details about the data from https://www.kaggle.com/mfaisalqureshi/hr-analytics-and-job-prediction

• Try to understand the different variables using some basic statistical measures and some visualization tools. As the minimum, calculate mean values for numeric columns, value counts for categorical column, correlation of the columns, and a couple of scatter-plots and box-plots to understand the relationship between the outcome column 'left and other variables'.

• Before creating classification models, transform all the categorical columns into numeric, and afterwards create traning and test sets (25% test set)

• Create a baseline model as a decision tree without any pruning (parameter specifications). What is the accuracy you can achieve?

• Create models by optimizing the parameters of (i) decision trees, (ii) bagging, and (iii) random forest classifiers. What is the best accuracy you can achieve across all the models?

• What are the three most important predictors according to the best random forest model? Create a new random forest model that uses only those three variables as predictors. What is the best accuracy you can achieve using only the three variables?

Task 02: Regression

In this exercise, you will perform tasks faced by a data scientist working in the real estate industry. Your job is to build a predictive model to estimate selling price for houses.

You can find more details about the dataset (Housing.csv), including an explanation of the variables from https://www.kaggle.com/mssmartypants/paris-housing-price-prediction

You need to perform the following tasks:-

- Exploratory data analysis: try to understand the different variables in the data. Identify the variables, based on exploratory data analysis methods (similar as in the previous task), that you think have an effect on the price of the house

- Develop a decision tree regression model that the company can use to predict the selling price for new houses on the market. Try to optimize the parameters. What is the best MSE that you can achieve?

- Create predictive models of neural network using keras. Experiment with different number of layers (up to 4). What is the best MSE that you can achieve?

Task 03: Text Analysis

In this exercise you will have to work with the data provided in the file ’text data.csv’. The data contains information on 165 articles from the San Francisco Chronicle from the Biz & Tech, Food, and US & World categories. You can find the following information:

- date: date when the article appeared online
- text: the content of the article
- title: the title of the article
- category: the category of the article

In the analysis, you have to work with the column text. Perform basic text preprocessing using stemming to obtain the most frequent words across all the articles. Try to iterate it at least two times, and in each iteration you should extend the set of stopwords with new ones based on the words you obtained as frequently occurring but are not particularly informative when you try to understand what the news is about. What are the 15 most frequent words after 2 iterations of removing various stopwords?

In the second step, perform topic modeling on the data with specifying three topics to be extracted. Based on looking at the top 15 words from each topic, can you identify a connection between the three original news categories and the three extracted topics?

Task 04: Predictive and Text Analysis

In this exercise you will have to you will have to analyze a dataset (new york airbnb.csv) that includes information about hosts, geographical availability, and different metrics available from Airbnb places in New York City. The list of available variables is as follows:

- id: listing ID
- name: listing Title
- host id: ID of Host
- host name: fame of Host
- neighbourhood group: name of location that contains listing
- neighbourhood: name of neighbourhood that listing is in
- latitude: latitude of listing
- longitude: longitude of listing
- room type: type of public space that is being offered
- price: price per night, USD
- minimum nights: minimum number of nights required to book listing
- number of reviews: total number of reviews that listing has accumulated
- last review: date in which listing was last rented
- reviews per month: total number of reviews divided by the number of months the listing is active
- calculated host listings count: amount of listing per host
- availability 365: number of days per year the listing is active
- Build a prediction model for predicting price with random forest regression (do not use ID columns, dates, and the name variable; additionally you may not want to use categorical variables with too many possible categories). Try to optimize the parameters of the model. What is the best result you can get?

- In the next step, you want to understand the impact of the review name on the price. First, identify the 15 most frequent words that appear in the listing column (after basic preprocessing, including changing to lowercase and stemming). Additionally to the general stopwords, think about what other expressions you need to exclude to get some meaningful results (e.g. you do not want things like ’york’, ’apartment’, ’room’, as they just refer to the specific location and context). Then add 15 new columns to the original dataset, with each new variable corresponding to one of the 15 identified frequent words, and the values of the new columns indicate whether that word appears in the name or not. For example, if you identify ’cheap’ as one of the most frequent words, you will need to create a new column, which will have value 1 in a row where ’cheap’ appears in the name, and 0 otherwise.

Finally, test whether the regression model that you created in the previous step can be improved byincluding these 15 new columns as predictors?