Assignment 1, Basic

Dennis Knol
2585404
dkl440

April 19, 2019

Task 1: Explore a Small Dataset

During the first lecture of the course, attending students created a dataset by answering questions. The goal of this exercise is to explore this data and to run simple experiments.

Task 1.A - Exploration

Exploring the data, we find that 276 students provided answers to sixteen questions presented by the lecturer. Of these questions, seven were multiple choice and nine were open questions. Initially, we focus on the answers to the multiple choice question. The key characteristic of multiple choice questions is that the possible answers are fixed, making it easy to quantify the answers and analyse the results. The answers to the multiple choice questions are presented in Figure 1.

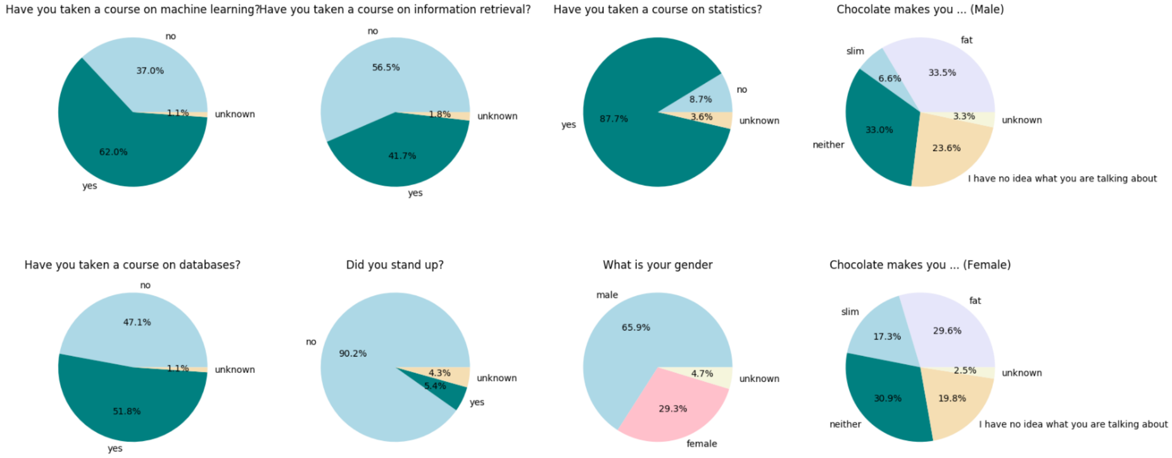

Figure 1: Pie plots presenting the answers to the multiple choice questions as a percentage

What we see from the charts, is the percentage of students that have taken a course on machine learning, information retrieval, statistics and databases. Most notably is that the vast majority of student have taken a course on statistics. Also, we note that only a relatively small number of students stood up and that the state “Chocolate makes you...” results in varied answers. This statement is also answered differently by men and women. For example, more women think chocolate makes you slim. The last multiple choice question for us to discuss is: “what is your gender?” An exploration of the results yield the following results: 66% of the students participating are men, 29% is female and 5% of the participant filled in the answer ‘unknown’.

Exploring the answers that were given to the open questions is a bit more challenging. For example abbreviations, different date formats and typos result in a diverse set of answers. Consequently, we first prepare the dataset.

Feature Engineering

We start by quantifying the multiple choice answers. For the Gender attribute, we set the value `male` equal to 0, `female` to 1 and `unknown` to two. We take a similar approach for questions which were answered with no, yes or unknown. These answers were converted to 0, 1 and 2 respectively. For the chocolate statement ‘fat’ is set to 0, ‘slim’ to 1, ‘neither’ to 2, ‘I have no idea what you are talking about’ to 3 and ‘unknown’ to 4.

Second is the preparation of the answers to the open questions. We start by categorising the programmes in which the participating students are enrolled. We convert abbreviations and programme names with typos into the written out name of the corresponding programme. Subsequently, we enumerated the programmes in alphabetical order (starting at 0 for AI, stopping at 19 for 'unknown'). Figure 2 presents the results by plotting the number of students per programme. The programme with the largest number of students is Artificial Intelligence.

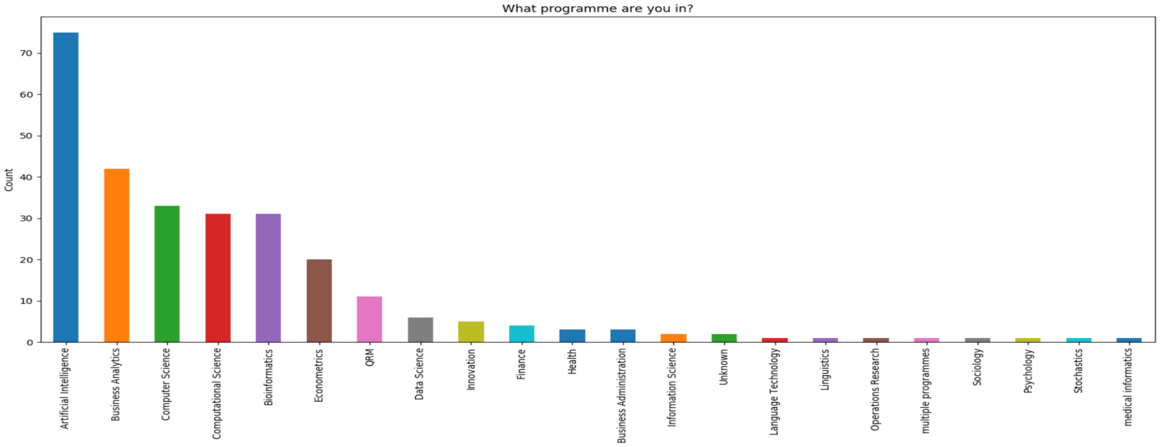

Figure 2: A count of the students in each programme.

We now discuss the answers to the question where we expect a numerical answer. These questions are the following: how many neighbours sit around you? How much money do you deserve? What is you're stress level? And lastly, give a random number. For all the corresponding features, we dropped the non numerical values and set limits.

For the features representing the number of neighbours we set a maximum value. This value is 342, as the capacity of the lecture room is 343.[1] The random number feature only considers values from 0 to 10 and the feature representing the amount of money someone thinks he/she deserves is limited to 100 euros. Lastly, all negative stress levels are set to 0 and all larger than 100 are set to 100. The values of the all four features are presented in Figure 3.

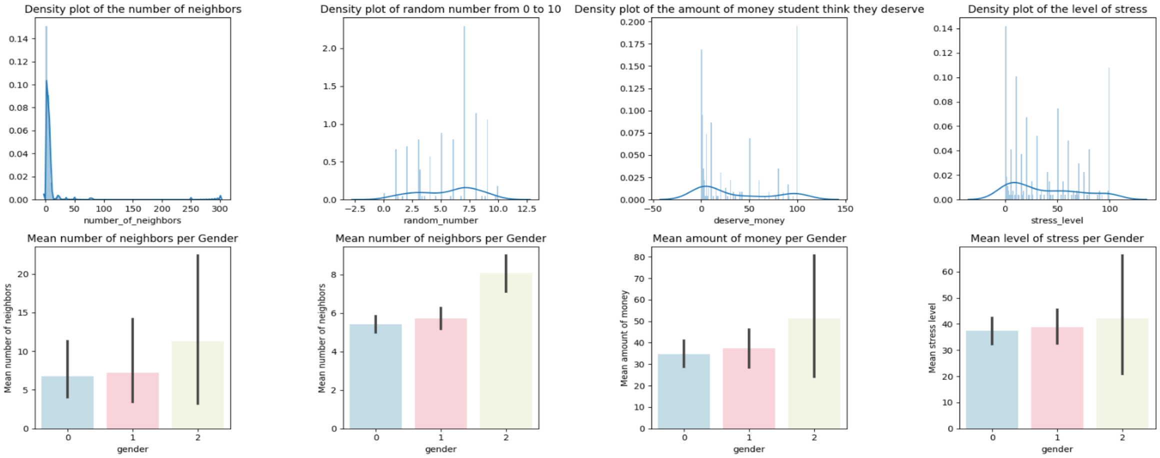

Figure 3: Density plots and bar plots for the number of neighbours, random number, deserved amount of money and stress level

The figure above shows the density plot, as well as a mean per gender. What we see in the density plots for the number of neighbours, is that the vast majority answered with a small number. The number occurring the most is 2 and the mean is equal to 7.1. For random number, the mode is 7 and the mean is 5.6. Looking at the plots on the bottom row, we see that the mean is highest for students with an unknown gender. Also, the difference between men and women is small.

Lastly, the answers to the questions “what makes a good day?”. We created various categories and categorised the answers. The categories creates are: sex, food, nice weather, alcohol, friends, sleep and sports. For each of these categories, we analysed what percentage is male or female. For this analysis we dropped the answers for which the corresponding gender is unknown. Figure 4 shows the results.

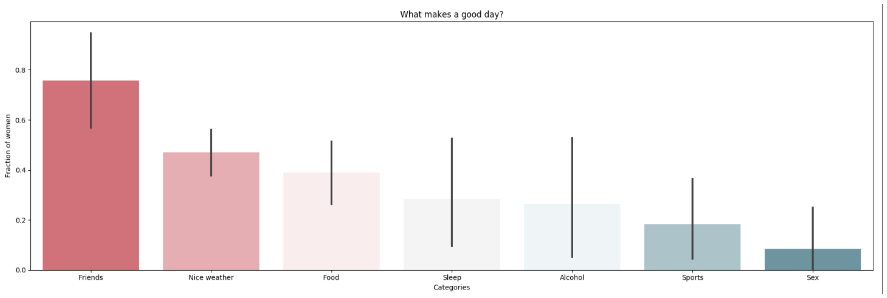

Figure 4: Gender distribution for answer to the question: what makes a good day?

The answers related to being with friends, are provided more often by female students when compared to male students. Answers related to nice weather are less dependent on the gender of the students. Close to 50 percent of the weather related answers are given by women. For the answers categorised in food, sleep, alcohol and sports we note that the fraction of men increases. Lastly, more than 90 percent of the answers related to sex are provided by male students.

Task 1.B - Basic Classification/Regression

We decided to to download a dataset from the web. The dataset is from the Machine Learning Repository and in called the Wine Quality Data Set.[2] What makes this particular dataset interesting to us, is that it combines a personal passion with our field of study.

The Machine Learning Repository included two data sets, one related to red and one related white wine samples. For our analysis, we focus on the red wine data and our goal is to predict the quality of the wine given the following attributes: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates and alcohol. The attributes are all non-null float values. The quality is a score from 0 to 10 and the values are integers. There are no values missing and thus there are 1599 values for each attribute.

To make is easier to predict the quality of the wine, we predict whether a wine is ‘good’ or ‘bad’. The good wines are the wine with a score higher than 6. Bad wines are the ones with a score lower or equal to six. The cross validation method we use is the train and test set method. The training set consists of two thirds of the data, the test set of one third. The results of three classification models are presented in the table below.

Table 1: Results from various classification algorithms

Classification Algorithm	**Accuracy**
Decision Tree classifier	88%
K-Nearest Neighbours classifier	88%
Random forest classifier	91%

Task 2: Predict Titanic Survival

The aim for this task is to predict whether or not a passenger survives, given the data provided by Kaggle. The competition provides both a train dataset as a test dataset. Only the train set contains information on an individual's survival. Consequently, we use this dataset set to train our models.

Task 2.A - Exploration and Preparation

To get an insight in the train data from the Titanic competition, we first examine what the attributes there are. Subsequently, we discuss the data types of each attribute and visually inspect the data. A preparation of the data follows the exploration.

The attributes of which the dataset consists, are the following: passenger ID, survived or not, class of travel, name of passenger, gender, age of passenger, number of siblings/spouse aboard, number of parent/child aboard, ticket, fare, cabin and embarked. The passenger ID is an integer and unique for each passenger. The survived attribute is equal to 1 for all passengers that survived and equal to 0 for all that didn't. The class in which a passenger travelled (first, second or third) is stored in attribute `Pclass`. The following three attributes contain the name, sex (male or female) and age of a passenger. Most of the age values are integers, some are floats. Also, we note that 177 Age values are missing. Next, the attributes `SibSP` and `Parch`. These contain integers representing the number of siblings or spouses aboard and the number of parents or children respectively. The Fare attribute contain float numbers representing the fair paid by a passenger. A passenger's ticket number is stored in `Ticket` and it's cabin number in `Cabin`. For the cabin number, we find that 687 values are missing. Lastly, the port in which a passenger embarked: Southampton, Cherbourg or Queenstown. This data is missing for two passengers. In what follows, we explore the attributes individually and discuss correlations.

Initially, we investigating the attributes Sex, Embarked and Class of travel. For each off the attributes we present a pie chart in Figure 5. We find that 64.8% of the passengers is male, 35.2% female and the majority of passengers, 72.4%, embarked in Southampton. Also, we note that 24.4% of the passengers travels in the first class, 20.7% in the second and 55.1% in the third.

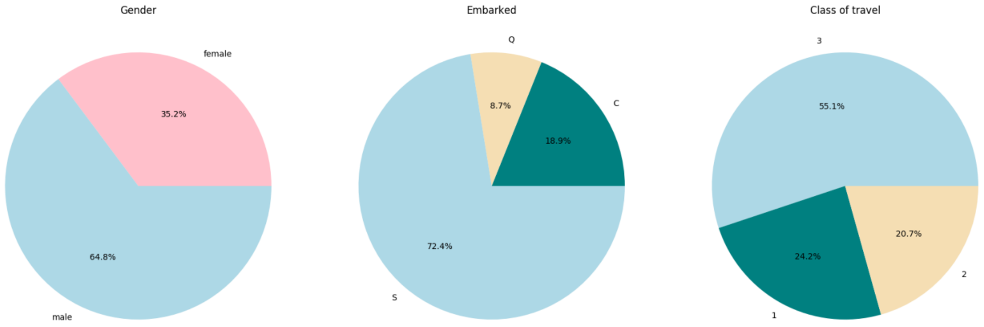

Figure 5: Pie chart of the attributes Gender, Embarked and Class of Travel.

A relevant finding form investigating gender, is that the survival rate for women is much higher. This if visualised in Figure 6. What we see is that of the chance of survival is must higher for women. This is most likely due to the evacuation policy of "women and children first". Other important findings are the correlation between survival and class of travel and survival and the port in which the passenger embarked. These results are also presented in the figure below. Also, as expected, we see a correlation between fare and class of travel. Fare increases from the third class to the first: mean fare in the third class is 13.86£, in the second 20.66£ and in the first 84.15£.

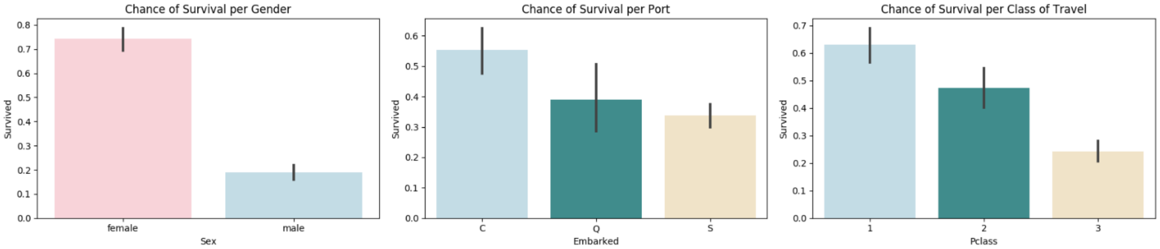

Figure 6: Chances of survival for gender, port of embarking and class of travel

We now explore the age attribute. In Figure 7 we see that that is a relatively small number of teenagers aboard and that majority of passengers is between 20 an 40 years old. What we did not find, is a clear relation between specific age and survived. However, when considering age categories we do see a correlation with survival. We have specified the following categories: babies (0-5 year old), children (6-15 year old), youth (15-24), adults (25-65) and seniors (65+). We find that babies are most likely to survive and seniors least likely. Also, we see that the average age increase from the third class to the first class: mean age in the third class is 25, in the second 30 and in the first 38.

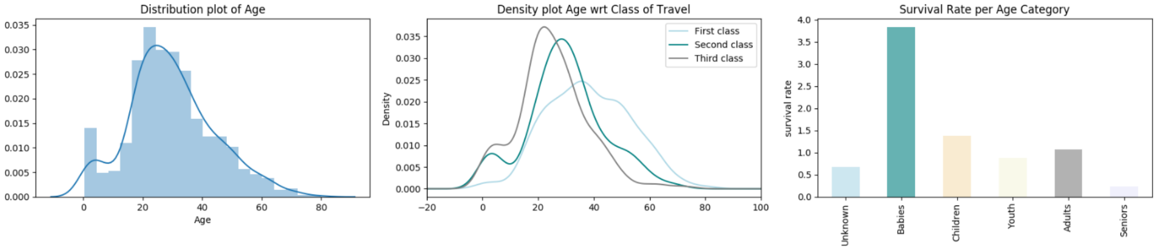

Figure 7: Density plots of age and of with respect to class of travel and the survival rate for each category

Last for the exploration and preparation part, is the detection and removal of outliers. For the outlier detection, we rely on the interquartile range (IQR) scores.[3] We drop passengers from the dataset when the value of more than two attributes is considered and outlier, which is the case for 11 passengers. Consequently, the train dataset now consists of 880 passengers.

Feature Engineering

Before we can apply various classifiers to the data, we have to convert the values of certain attributes. This selection describes the changes we made.

First the gender attribute. In the provided dataset, the gender of a passenger is written out into male or female. We convert the value `male` to 0 and the value `female` to 1. Similarly, we convert the letter which represents the port in which a passenger embarked. We set the Southampton port equal to 1, the Cherbourg port equal to 2 and the port in Queenstown to 3. The missing values were filled with the most appearing value: Southampton.

The Fare attribute consists of a larger number of unique values. Because of this reason, we sort all the values in 10 bins. The bins are numbered from 1 to 10. Also, we fill the missing values with the mean fare of the corresponding class of travel.

The attribute cabin has the most missing values. In the train set 687 values out of 891 in total are missing. These empty cells were filled with the value 0. From the known cabin numbers, we subtract the letter in front. This letter is relevant, as it contains information on where on the ship the cabin was located. We enumerated the subtracted letters from 1 to 8.

Lastly the attributes name and age. What we find in the values of the name attribute is not only the name of a passenger, but also his or her title. The title of a passenger is relevant, as it is highly correlated to survival and age. Consequently, we subtract this title from the values of the name attribute and save it in a new column. We conjoin titles that occur rarely. The following titles “Dr”, “Rev”, “Col”, “Major”, “Jonkheer”, “Don”, “the Countess”, “Lady” and “Sir” have been renamed to “Rare”. The column containing the names of the passengers is dropped.

The correlation between title is exemplified in Figure ... . The figure presents the age density for the titles Mr, Miss, Mrs and and the collection of rare titles. We see that the average age for unmarried women is 22 years and for married women 39 years. Men with the title “Mr” are on average 32 years old and passengers with rare title 45 years old. These results and the average age of the other titles are utilised to fill the missing age values. Lastly, we sort age in the categories babies, children, youth, adults and seniors.

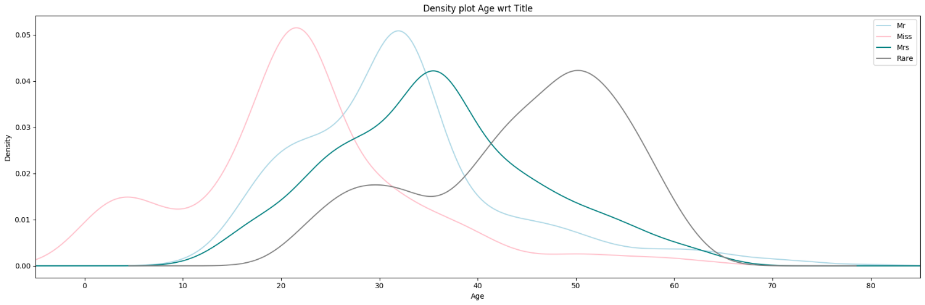

Figure 8: Density plot of age with respect to titles Mr, Miss, Mrs and and the collection of rare titles

Task 2.B - Classification and Evaluation

We now have an intuitive understanding of the training data. Also, the data is prepared and ready for evaluation and classification.

To start, we determine the explanatory variables that eventually predict whether a passenger survives. These variables are the class of travel, gender, number of siblings or spouses aboard, number of parents or children aboard, cabin, embarked, fare, age categories and title. Thereafter, we create a train set and a test set. The training set consists of two thirds of the data and the set set of one thirds of the data. Now that the appropriate setup is created, we apply and evaluate at various classification algorithms. We present the results in the table below.

Classification Algorithm	**Accuracy**
Decision Tree classifier	80%
K-Nearest Neighbours classifier	81%
Gradient Boosting for classification	79%
C-Support Vector classification	79%
Multi-layer Perceptron classifier	78%
Random forest classifier	84%

Table 2: Results from various classification algorithms

From the table, we see that the random forest classifier has the highest prediction accuracy and we thus apply this classifier to the test set from Kaggle. We submitted the resulting submission file and got a prediction score of 0.79904. The score got on in the top 15% of the competition.[4]

Task 3: Research and Theory

Task 3.A - Research: State of the Art Solutions

- Find a data mining competition that is already finished
- Description of the competition
- Who was the winner and what technique did they use
- What was the main idea of the winning approach
- What makes the winning approach stand out?

Task 3.B - Theory: MSE verse MAE

Both the mean squared error as the mean absolute error measure the mean model prediction error by considering the actual values, y_i, and subtracting the predicted value, $\hat{y}_i$. Both measures only output positive scores. The range is from zero to infinity.

What differentiates the two, is the term which is averaged. The MSE squares the errors and the MAE takes the absolute value. The corresponding formulae are presented in the following. The first formula represents the mean squared error. The one second is the formula for the mean absolute error. In both formulae, Y_i a vector of the actual values and $\hat{Y}_i$ a vector of the predictions.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_I)^2 \quad \text{and} \quad MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

The difference between the two measures is most significant when the data that is being analysed contains outliers. Large error are weighed heavily by the MSE as a result of squaring the errors. Also, the MSE tends to be larger for increasing sample sizes. To conclude, the MAE should be used over the other.

Mathematically, the equations are exactly the the same when $y_i = \hat{y}_i$ or $|y_i - \hat{y}_i| = 1 \ \forall i$. An example of such situation is discusses in what follows.

Experiment

To exemplify the behaviour of the specific error measures discussed above, we run an experiment on a dataset obtained from the web.[5] The dataset consists of data of 365 consecutive days, covering entire 2015, and of 7 attributes. For the regression, we consider consumption as the dependent variable. The explanatory variables are mean, minimum and maximum temperature, precipitation and a dummy variable for weekend.

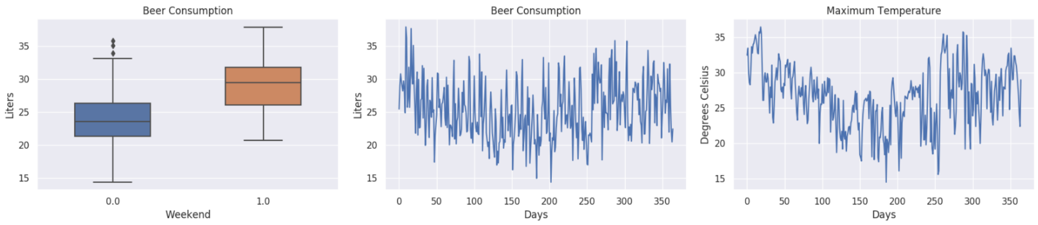

Figure 9: Exploration plots

Using Figure 9 we visually explore the data. What we see in the boxplot is that the consumption of beer is higher during the weekend. From the two plots on the right hand side, we see that the consumption of beer is lower when the temperature is lower, indicating a direct relationship between temperature and the consumption of beer.

We use these finding from the data to draw up to experiments. The first is a linear regression with explanatory variables precipitation, weekend and maximum temperature. Using these variables, we predict the consumption of beer. The second is a logistic regression predicting whether it is weekend, using the variables consumption and mean temperature. The distribution off the errors in both experiments is presented in the figure on the next page, Figure 10.

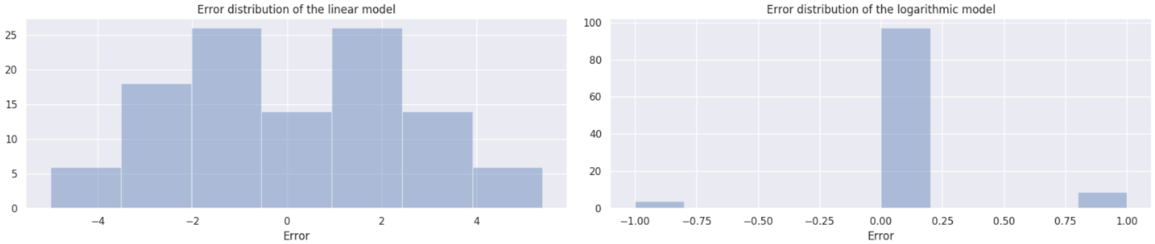

Figure 10: Histograms of the errors from the linear and logarithmic model

What we see from the figure is that the errors from the linear regression are more dispersed than the errors for the logistic regression. In other words, the errors of the logistic regression are more evenly distributed. Consequently, the MSE is larger than the MAE for the linear regression. For the logistic regression, we see that the MAE and MSE are identical (when rounded to 4 decimals). This is because $y_i = \hat{y_i}$ or $|y_i - \hat{y_i}| = 1$ $\forall i$. The results are tabulated below.

Classification Algorithm	**Accuracy**
MSE of linear regression	5.5834
MAE of linear regression	2.0277
MSE of logistic regression	0.1182
MAE of logistic regression	0.1182

Table 3: Results from various classification algorithms

Task 3.C - Theory; Analyse a Less Obvious dataset

The aim of this exercise is to predict a message to be spam or ham. A modelling technique that is suitable for the text message data, is for example the Naive Bayes classifier. We will use the TF-IDF vectorizer. The vectorizer transforms text to feature vectors that can be used as input to estimator.

The dataset consists of two columns. The first column is for the label, the second is for the the text. We start by converting ham to 0 and spam to 1. Quickly exploring the data, we find that 86.59% of the messages are labelled as ham and 13.41% is labelled as spam. We also find that there two missing values in the text column. We drop the corresponding instances. Lastly, we transform the text data by converting uppercase to lowercase.

After making some changes to the data, we split the data into a test and a train set and vectorize the text. Subsequently, we train the vectorizer and use Support Vector Machine classifier. This results in an prediction accuracy score of 83.25%. Considering the simplicity of our approach, we think there is room for improvement. For example, the data processed a further (e.g. stop words could be removed) and we could use a different classifier (e.g. Naive Bayes classifier).

References

[1] VU, URL: `https://www.vu.nl/nl/Images/Zaalfaciliteiten_aug2018_tcm289-257362.pdf`

[2] Machine Learning Repository, URL: `https://archive.ics.uci.edu/ml/datasets/wine+quality`

[3] Towards Data Science, URL: `https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba`

[4] Kaggle, URL: `https://www.kaggle.com/dennisrkn`

[5] Kaggle, URL: `https://www.kaggle.com/dongeorge/beer-consumption-sao-paulo`