Hochschule München University of Applied Sciences

PROJECT WORK: AIRLINE CUSTOMER ANALYTICS

M4.04 - Big Data and Artificial Intelligence

Please grade individually as marked chapter-wise.

Anna-Christine Margot Saam - Stefanie Emmerling – Muhammad Ahmed

Table of Contents

1. Introduction (Anna-Christine Margot Saam & Stefanie Emmerling)

This report documents the evaluation of the airline passenger satisfaction survey. The dataset that is analysed in this report has been provided by the client. The analysis of this dataset is realized using different Python packages such as pandas, numpy, matplotlib, sklearn and seaborn. These packages provide functions that have been built for data analysis. The Jupyter Notebook containing the full code is provided additionally to this report, the link is available for your reference in the Appendix.

To start with, this report will give an insight on the executed data cleaning that is necessary to prepare the data for the analysis of different aspects of the data. Furthermore, there will be a chapter on customer satisfaction and what factors can be used to predict a satisfied passenger and giving recommendation on how that might be helpful for the airline to develop strategies to improve the overall satisfaction of their customers. To conclude, there will be a chapter on the characterization of different groups of customers and how this information can be used in the airline's operations.

2. Data Preparation (Anna-Christine Margot Saam & Stefanie Emmerling)

The head of the provided dataset is displayed below, for formatting reasons split in two images.

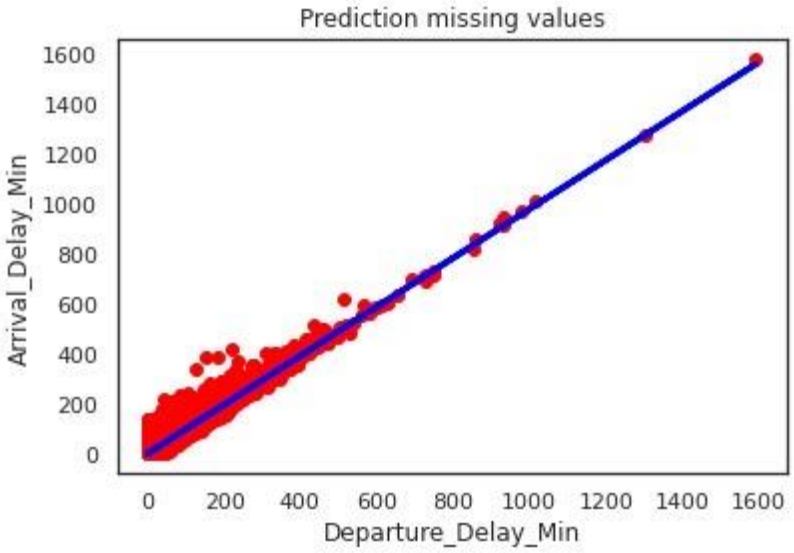
	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink
0	0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	3	1	5
1	1	5047	Male	disloyal Customer	25	Business travel	Business	235	3	2	3	3	1
2	2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	2	2	5
3	3	24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	5	5	2
4	4	119299	Male	Loyal Customer	61	Business travel	Business	214	3	3	3	3	4

Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	satisfaction
3	5	5	4	3	4	4	5	5	25	18.0	neutral or dissatisfied
3	1	1	1	5	3	1	4	1	1	6.0	neutral or dissatisfied
5	5	5	4	3	4	4	4	5	0	0.0	satisfied
2	2	2	2	5	3	1	4	2	11	9.0	neutral or dissatisfied
5	5	3	3	4	4	3	3	3	0	0.0	satisfied

After the first check-up the following general information on the dataset could be retrieved: The dataset contains 25 columns with a total of 103924 entries. The data type of all objects is correct, meaning numerical values are not falsely displayed as an object and would be interpreted incorrectly in the further analysis. There is just one column that has 310 missing values, which is the column that is giving information on the arrival delay.

2.1 Dealing with empty values

Rather than deleting these 310 rows, that contained missing values in just one column, it was decided that these missing values for the arrival delay should be predicted with the help of a linear regression and the values of the departure delay column. The following chart shows the linear regression.



The graph is nearly representing the linear function $y = f(x) = x$. This means that the values correlate almost in a one-to-one relationship. So, for example, if the flight departed 45 minutes late, an arrival delay of 44.83 minutes is expected.

The replacement of the missing values is of minor importance due to the low number of lines affected in comparison to the total data set. But for future use cases this can be of great importance and is therefore demonstrated.

2.2 Unify data

A screening of the data for unique values was performed as different issue are common with raw data sets such as spelling of the same response in different ways. This can cause falsified outcome of the analysis. It was identified that lower- and upper-case spelling in column Disloyal_Customer and Business_Travel differs. These unique values were harmonized and

saved in the newly created df_clean data frame. In addition, it surfaced that the column ID contains no duplicate values. Hence it cannot be interpreted as a customer ID but probably reflects a survey ID, being unique for each line of the data set. Finally, the column Baggage Handling only received a rating ranging from 1 to 5 but no 0 rating. It is rather unlikely that all customers replied above 0 as 0 was defined as not applicable. However, having no further option to investigate why this issue occurred, in the following analysis the column will remain as it was delivered in the raw data set.

2.3 Dealing with no answers and minorities

Analysing further columns with the unique value analysis to execute further data cleaning the following tables show the amount of different single values.

The column „Satisfaction“ contains three unique entries:

- neutral or dissatisfied
- satisfied
- no answer

The column „Gender“ contains 4 unique entries:

- Male
- Female
- Na
- Diverse

For the further analysis of satisfaction: "no answer" and Gender "na" is not value adding. As the number of lines containing these values is respectively 6 and 9, these lines were removed. Furthermore only 7 lines contain the value „diverse“. Since this is not significant for the total data set these lines will also be dropped. The information loss from the other columns is not significant. However, if further analysis on diverse gender is required and relevant for the company, more data from diverse clients is needed.

2.4 Dummyize categorial variables

The dataset contains several features that are categorial: Gender, Customer_Type, Type_of_Travel, Class and Satisfaction. To realize clearer data analysis these features will be converted to dummy variables, which is a binary variable corresponding to one value of a categorial variable. The above-mentioned variables will appear slightly different in the

following report, as they are split into their different classes. Satisfaction for example will appear as Satisfaction_neutral_or_dissatisfied and Satisfaction_satisfied.

2.5 Standardize and revise response scale

Most columns contain data that is referring to a response from the survey meaning "not applicable". The dataset is reflecting 0 values in the following columns:

- Inflight_Wifi_service
- Departure_Arrival_Time_Convenient
- Ease_Of_Online_Booking
- Gate_Location
- Food_And_drink
- Online_Boarding
- Seat_Comfort
- Inflight_Entertainment
- On_board_Service
- Leg_Room_Service
- Checkin_Service
- Inflight_Service
- Cleanliness

For these columns the mean value of the rest of the participants response will be set in place of the 0 values, so that it will be possible to use these lines for different models. Keeping the 0 values, would put those responses in a classification as the lowest, meaning the least satisfied result. However, that is not what it was meant to mean.

3 Customer Satisfaction (Anna-Christine Margot Saam & Stefanie Emmerling)

The following chapter is focusing on an analysis of the customer satisfaction and whether there are variables that can predict customer satisfaction. If so, these variables will be pointed out. Furthermore, according to these variables, recommendations are given on how this information can be used to improve the overall satisfaction.

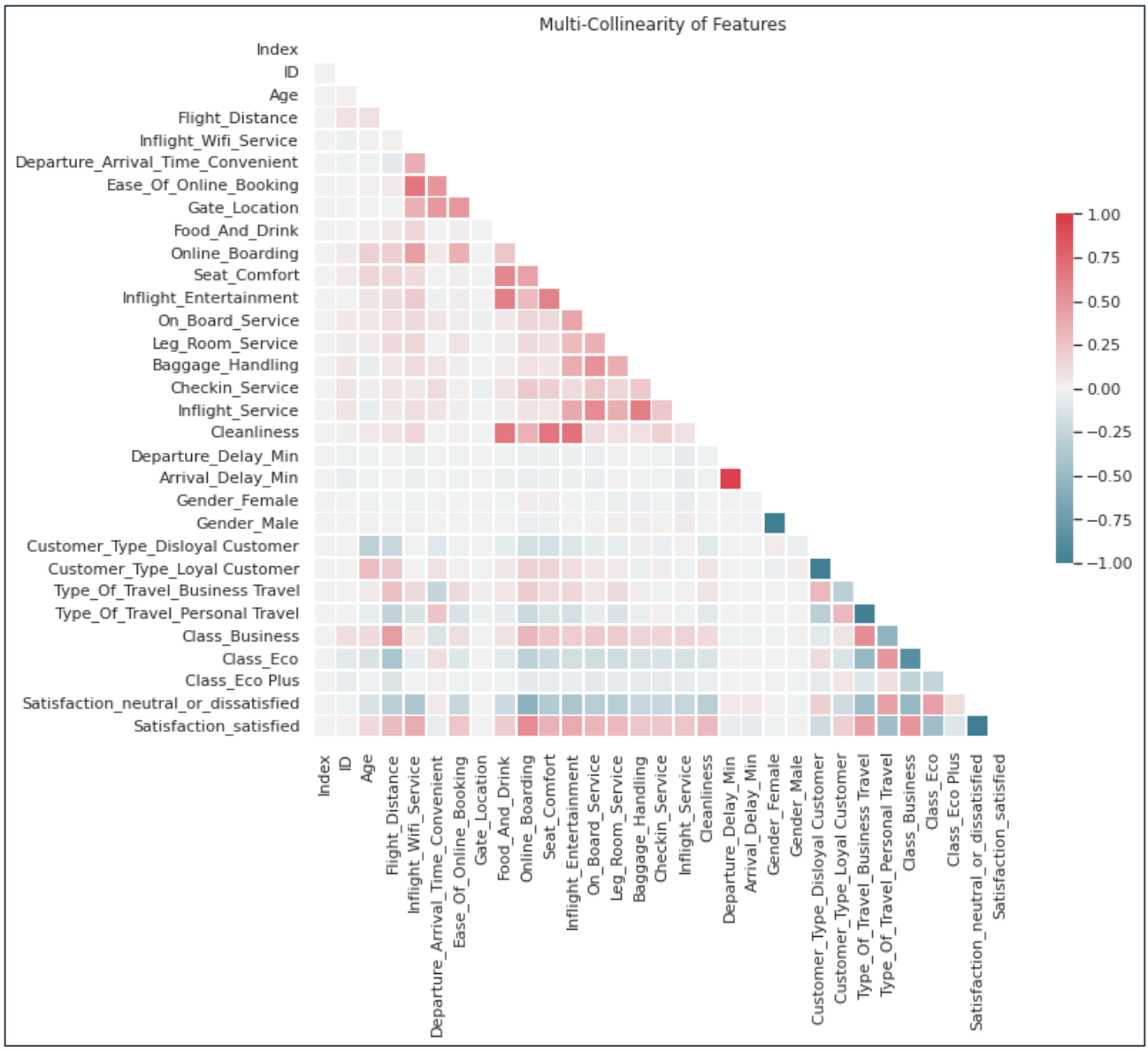
3.1 Relevant factors for satisfaction level of passenger

This section will focus on the relevant factors for satisfaction level of the passengers of this airline. As a first step the factors correlating with satisfaction will be identified and later the

relevant factors will be modelled with a decision tree, that will help to predict customers satisfaction very intuitively.

3.1.1 Correlation of relevant factors

To apply a data model, it needs to be checked which features correlate with the dependent variable. The dependent variable is Satisfaction as the aim is to find out what other features influence Satisfaction. Among the independent variables those independent features need to be eliminated that are correlating as they are giving redundant information. Correlating independent variables influence the accuracy of the results of the model negatively. An appropriate way to check this is using a heat map, which also gives a good visualization of the math behind it. The darker the colour the higher the correlation, while red indicates a positive correlation and blue a negative correlation between the respective variables.



The heat map shows that values that correlate negatively with Satisfaction_satisfied correlate positively with Satisfaction_neutral_or_dissatisfied and vice versa. When displaying the numerical value of each variable combination, this visual impression can be proven in numbers. R expresses how strong the correlation is, ranging from 0, which is no correlation, to 1, which is highly correlating. The absolute R values for the independent variables in correlation with satisfaction are the same for both satisfied and neutral or dissatisfied, as can be seen in the following table.

	Satisfaction satisfied	**Satisfaction neutral or dissatisfied**
Satisfaction_neutral_or_dissatisfied	-1.000000	1.000000
Class_Eco	-0.451147	0.451147
Type_Of_Travel_Personal Travel	-0.448950	0.448950
Customer_Type_Disloyal Customer	-0.187666	0.187666
Class_Eco Plus	-0.105306	0.105306
Arrival_Delay_Min	-0.057645	0.057645
Departure_Delay_Min	-0.050546	0.050546
Departure_Arrival_Time_Convenient	-0.047953	0.047953
Gender_Female	-0.012294	0.012294
Gate_Location	0.000727	-0.000727
Gender_Male	0.012294	-0.012294
ID	0.013824	-0.013824
Age	0.137148	-0.137148
Customer_Type_Loyal Customer	0.187666	-0.187666
Food_And_Drink	0.210734	-0.210734
Ease_Of_Online_Booking	0.234721	-0.234721
Checkin_Service	0.236138	-0.236138
Inflight_Service	0.244675	-0.244675
Baggage_Handling	0.247737	-0.247737
Flight_Distance	0.298771	-0.298771
Cleanliness	0.305049	-0.305049
Leg_Room_Service	0.315944	-0.315944
On_Board_Service	0.322340	-0.322340

Seat_Comfort	0.349460	-0.349460
Inflight_Wifi_Service	0.381964	-0.381964
Inflight_Entertainment	0.397951	-0.397951
Type_Of_Travel_Business Travel	0.448950	-0.448950
Class_Business	0.503841	-0.503841
Online_Boarding	0.557976	-0.557976
Satisfaction_satisfied	1.000000	-1.000000

This behaviour implies that the data is correct from a logical perspective.

Since satisfied and dissatisfied are two sides of the same medal the analysis will continue with the Satisfaction_satisfied side. A value near to 0 (both positive and negative) indicates the absence of any correlation between two features, which means these variables are independent of each other. It will not be useful to include them in a model since they have simply no influence on Satisfaction.

The following features are (highly) influencing the independent variable Satisfaction_satisfied and are also printed bold in the table (listed descending per absolute R score, cut was made at 0.29 as there is then a small gap to 0.24).

- Online_Boarding
- Class_Business
- Class_Eco
- Type_Of_Travel_Business Travel
- Type_Of_Travel_Personal Travel
- Inflight_Entertainment
- Inflight_Wifi_Service
- Seat_Comfort
- On_Board_Service
- Leg_Room_Service
- Cleanliness
- Flight_Distance

Now the inter-correlation of the independent variables will be analysed looking at the highest-ranking variable Online_Boarding. In theory this means the higher a customer rates

Online_Boarding the higher will be the score for the intercorrelating variable or vice versa should the correlation be negative. Visually in the heat map, this can be seen by checking the column and row combination of Online_Boarding and the other variable in question.

The following list shows the intercorrelated variables that will be dropped due to high R score, that would, as mentioned before, only give redundant information:

- Seat_Comfort 0.433368
- Inflight_Wifi_Service 0.457645
- Class_Business 0.324313
- Class_Eco -0.288713
- Inflight_Entertainment 0.293615
- Cleanliness 0.346841

The second list of variables shows those with low R score and will be kept together with Online_Boarding:

- Type_Of_Travel_Business Travel 0.209975 (correlated with Type_Of_Travel_Personal Travel)
- Type_Of_Travel_Personal Travel -0.209975 (correlated with Type_Of_Travel_Business Travel)
- On_Board_Service (correlated with leg room service) 0.166337
- Leg_Room_Service (correlated with on board service) 0.131906
- Flight_Distance 0.194582

The information derived from the visual check of the heap map unfolds that the set above has pairs of variables that are not highly correlating with Online_boarding but with each other. This information is stated in brackets. Again, keeping all variables would create redundancy and not result in a better model, therefore only one of each pair is used. The decision was to keep the ones with the highest positive R score.

3.1.2 Model Selection

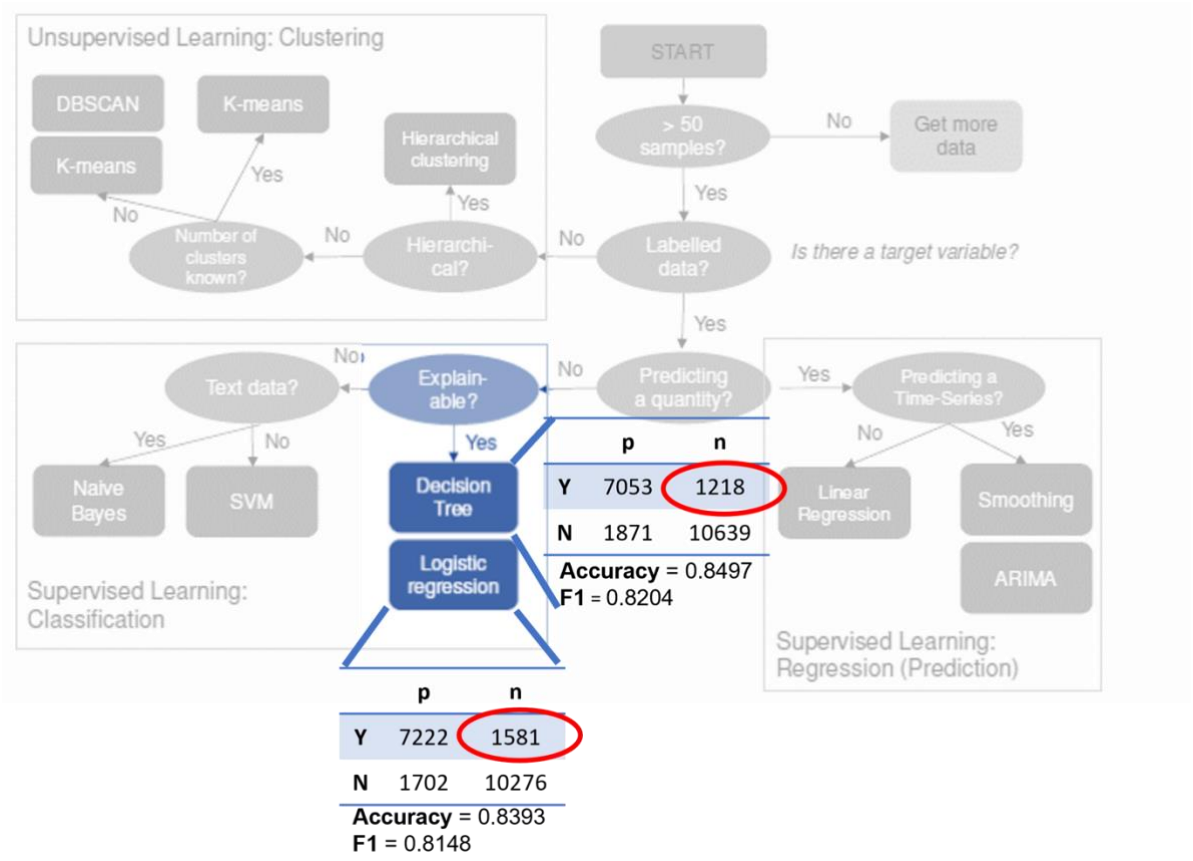
This section will discuss and explain the selection of the model that has been used to analyse the dataset. To start with, the final selection of variables will be explained. Not only the model itself, but also the choice of variables that will be included in the model make a difference. As mentioned before, keeping intercorrelating variables, is not value adding. But more than that,

keeping these variables could also causing overfitting as the model is trained towards adjusting the categories to outliers and increasing the probability of the model resulting in a false decision. Therefore, several combinations of variables, with the two different models have been tried out. The following code, can also be found in the Jupyter Notebook, shows the three different options that were used to check the different models for their accuracy, confusion matrix and f1 score.

```
# Create feature vector and success variable
#Option 1 (all that correlate >.29 without excluding intercorrelation
of independt values)
#X = df_clean[['Online_Boarding','Class_Eco','Class_Business',
'Type_Of_Travel_Business Travel','Type_Of_Travel_Personal Travel',
'Inflight_Entertainment','Inflight_Wifi_Service','Seat_Comfort',
'On_Board_Service','Leg_Room_Service','Cleanliness','Flight_Distance']]
#Option 2 (>.29 and intercorrelation excluded)
#X = df_clean[['Online_Boarding','Type_Of_Travel_Business Travel',
'On_Board_Service','Flight_Distance']]
#Option 3 (>.29 and intercorrelation excluded)
X = df_clean[['Online_Boarding','Type_Of_Travel_Business Travel',
'Leg_Room_Service','Flight_Distance']]
Y = df_clean['Satisfaction_satisfied']
```

As provided in the Jupyter Notebook, if of interest, it is possible to rebuild the models with those other options as well. However, for further analysis it was decided, to go ahead with Option 3, keeping Online_Boarding, Type_Of_Travel_Business Travel, Leg_Room_Service and Flight_Distance as the X features to build the model. The final selection was based on the correlation check as discussed in the previous chapter.

There are different models that can be used to further analyse the factors that have an impact on customer satisfaction. As can be seen in the following figure there are basically two models that need to be considered. The present data set needs a classification for the customers of the airline, and it is furthermore explainable data (meaning there are no data classes such as yes or no). Therefore, the possible options are a Decision Tree or Logistic regression.



Source: (Anderl, 2021)

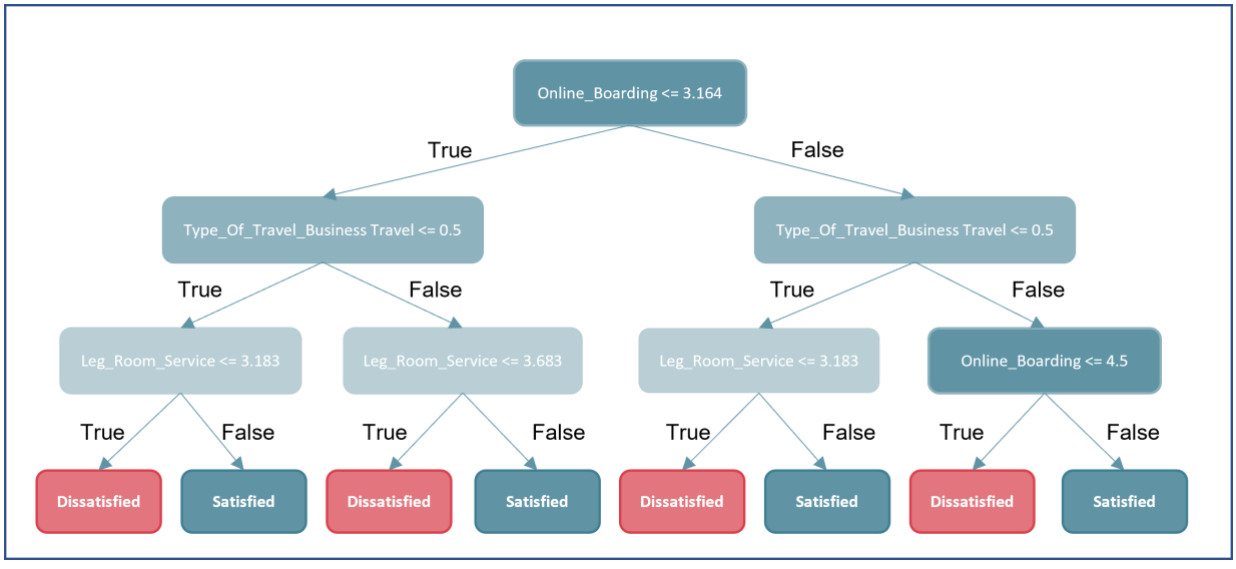
To train the models it is important to split the dataset in train and test data. Therefore, they have been separated with a test-train split of 20/80. Using the test data, the accuracy of the models was detected, and further evaluation has been done. The two models have already been considered to decide on the variable selection. For the final decision on the model however, there were two main reasons, that made the difference. First, the accuracy, f1 and the confusion matrix have better results for the decision tree. Furthermore, the decision tree is way more intuitive to read and understand and makes it therefore the better choice to work with especially when business implication is needed to derive from it.

To summarize the main evaluation outcomes of the decision tree, the calculated metrics will be explained in more detail. The confusion matrix of the decision tree can be interpreted as an overview of the risk of wrong interpretation of the data. The red circled values in the figure are the false positives, being the number of customers, that would wrongly be interpreted as satisfied customers, whereas they are neutral or dissatisfied customers in reality. For this present practical use, this can be seen as the worst mistake that can be done, which is why it should be reduced to a minimum. As the decision tree has a lower number in the metric,

compared to the Logistic regression, it affirms the choice of the decision tree as the model to continue with. The f1 score can be interpreted as the weighted average of the precision and recall values. Since the f1 score is reaching a maximum of 1 as its best outcome and a minimum of 0 as its worst, reaching 0.8204 is very good already, being just a bit better than the logistic regression metrics again. This is the same also for the accuracy of the decision tree, which describes how good the model is describing the data.

3.1.3 Implementation and interpretation of the Decision Tree

The final step then is to implement the Decision Tree, using the training data, as separated before. The training data is always used to train the model. If combinations of other variables or a tree with a higher depth, meaning more layers than the current three below, are of interest, the Jupyter Notebook includes a section where the visualization of the decision tree is directly updated when changing the variables and rerunning the code. In this report however, the focus will still be on the selected variables as explained in the chapters above. The depth will also be kept at three as adding more layers does not improve the combination of accuracy, confusion matrix and f1 score significantly and decreases the ease of application. In addition, to be able to get business relevant information out of this model, this chapter will conclude with an interpretation of the decision tree. The following decision tree is the outcome of training the model with the carefully prepared dataset.



As mentioned before, this section will also include a first interpretation of this decision tree. Answering the main question, whether a customer will be satisfied or not, can be done by looking at the final row. To get a clearer picture within the decision tree, the variable Satisfaction_neutral or dissatisfied has been shortened to ”dissatisfied”, however including

the same data. Different categories that are influencing the satisfaction of the customer, can be seen in the first three levels of the tree. The highest impact, as could already be seen in the heat map, has the variable Online_Boarding. Furthermore, the type of travel appears to be important. The variable Type_Of_Travel_Business Travel can only be 0 or 1, 0 meaning false, being a customer that selected Personal Travel as type of travel and 1 meaning true, meaning a customer that chose business travel as type of travel.

To make an example, one path along the branches will be outlined in more detail, starting at the top of the tree and following the branches always to the right side. One fictive customer, that is using the airline for Business Travel will for example be satisfied, according to the model, when rating the Online_Boarding with 5.

1. Leaf: Online_Boarding 5.0 ≤ 3.164 FALSE
2. Leaf: Type_Of_Travel_Business Travel 1 ≤ 0.5 FALSE
3. Leaf: Online_Boarding 5.0 ≤ 4.5 FALSE

→ Satisfied

As it is not possible to outline every single possible way of going through the model, this was just one example on how it can be used. However, the following chapter will provide further information and recommendations on how to use the findings of the model to improve customer satisfaction.

3.2 Recommendations to improve overall satisfaction

The results of the analysis give valuable information on how to improve customer satisfaction. Besides the variables used in the model, from a business perspective all variables that positively correlate with Satisfaction can be of use to improve the business model of the airline. Therefore Online_Boarding, Class_Business, Type_Of_Travel_Business Travel, Inflight_Entertainment, Inflight_Wifi_Service, Seat_Comfort, On_Board_Service, Leg_Room_Service, Cleanliness, Flight_Distance should be analysed. Without further knowledge on the detailed operations and cost implications the following advice is given:

Online Boarding was the variable that correlated the strongest with a satisfied customer. Therefore, improving online services is strongly recommended. In order to save time and create more convenience user accounts for easy check in should be made available. The customers can then save their personal information such as passport number without having

to re-enter the information each check in. Furthermore, the available online options can be advertised to the customer by sending email reminders. When online boarding has opened sending an email is a low effort option for the customer to use this service. Applications for Android and iOS are an add on for additional convenience and usability as mobile website traffic accounts for approximately half of web traffic worldwide (Statista 2022).

In relation to the flight distance, it surfaces that the longer the distance the more satisfied a customer is. Therefore, it is recommended to implement certain services that are rather likely for long distance flights on shorter routes as well. These services can be snacks, inflight entertainment, Wi-Fi service, on board service or possibly revised baggage options. This is also reflected in the correlation of the variables Inflight_Entertainment, Inflight_Wifi_Services, On_Board_Services in relation to Satisfaction. From an economical perspective providing the full extent of services as available on long distance flights might not be feasible. However, low-cost options like offering cookies and a coffee instead of a full breakfast should be considered.

Customers that fly for business reasons are more satisfied. The expectations for a flight are different. This type of customer often works on flights, not prioritizing for example available entertainment options. For them on demand services are more important than frequent interruptions by flight attendants. In addition, it is suggested to analyse route options and their respective flight times so that the available schedules are more suitable for business customers. This will in return attract more business customers in comparison to competitors.

Comparing the different classes, business class customers are more satisfied as by nature business class offers better services in the with Satisfaction positively correlating categories of Inflight_Entertainment, Inflight_Wifi_Service, Seat_Comfort, On_Board_Service, Leg_Room_Service. An option to improve satisfaction of customers in eco class is the possibility to upgrade to better leg room service at a reasonable price. Typical examples here are the first-row seats and seats located at the emergency exits. Another possibility is to offer last minute upgrades via e-mail to a better class if the occupancy rate of the flight is low. Another option to provide upgrades at a low cost is the introduction of a customer loyalty programme. The collected points can be used on upgrades if the better classes have availability. In addition to a higher satisfaction rate this also binds the customer to the airline.

The above analysis was provided on the information available and is in essence a first collection of ideas. Further recommendations can be made if detailed insights on the current operations are made available. In addition, clarification on the dataset provided would be required. During the data preparation phase several adjustments such as replacing certain values with the mean value were performed with limited information on the raw data. This process is a potential source for errors and can influence the result of the data evaluation.

4 Groups of similar customers (Muhammad Ahmed)

In this part we will try to group customers into clusters using their survey responses and the data available. The main crux of this chapter will be to identify customer segments for the airline using a data mining approach. The characteristics of these customer segments will then further be used to understand how to serve the customers better and what the airline can do to increase the satisfaction of their customers.

This customer segmentation is important because of two core reasons. Firstly, it will help the airline understand how some customer groups have similar or different preferences. Secondly, it will make the task of addressing customer concerns and increasing satisfaction of the overall customers, much easier than if the airline did this analysis on an individual level due to there being too many customers that such a task would become impractical.

In order to group our customers, we will use clustering to deduce different customer segments. A cluster is a group with similar data objects such as observations, patterns and units that can be grouped together by virtue of being similar to each other while at the same time being unique from other clusters or groups. (Xu and Wunsch, 2008) Cluster analysis categorizes the dataset into a certain number of sets or groups whereby each set has elements that appear similar but appear dissimilar to elements of other sets or groups.

4.1 K-means Clustering

We will do the clustering based on the insights gained from the chapters two and three. We saw that Online-Boarding was the biggest factor in influencing a positive satisfaction score in the survey. We also found out that many other variables were very closely linked to Online_Boarding which is why adding them to the data frame will not add much value here.

Which is why we will only choose variables which had a low inter-correlation score with Online_Boarding, which as we saw earlier are: Type_Of_Travel_Business Travel, Type_Of_Travel_Personal Travel, On_Board_Service, Leg_Room_Service and Flight_Distance. Hence our X is:

```
X = df_clean[['Online_Boarding','Type_Of_Travel_Business Travel','Type_Of_Travel_Personal Travel','On_Board_Service','Leg_Room_Service','Flight_Distance']]
```

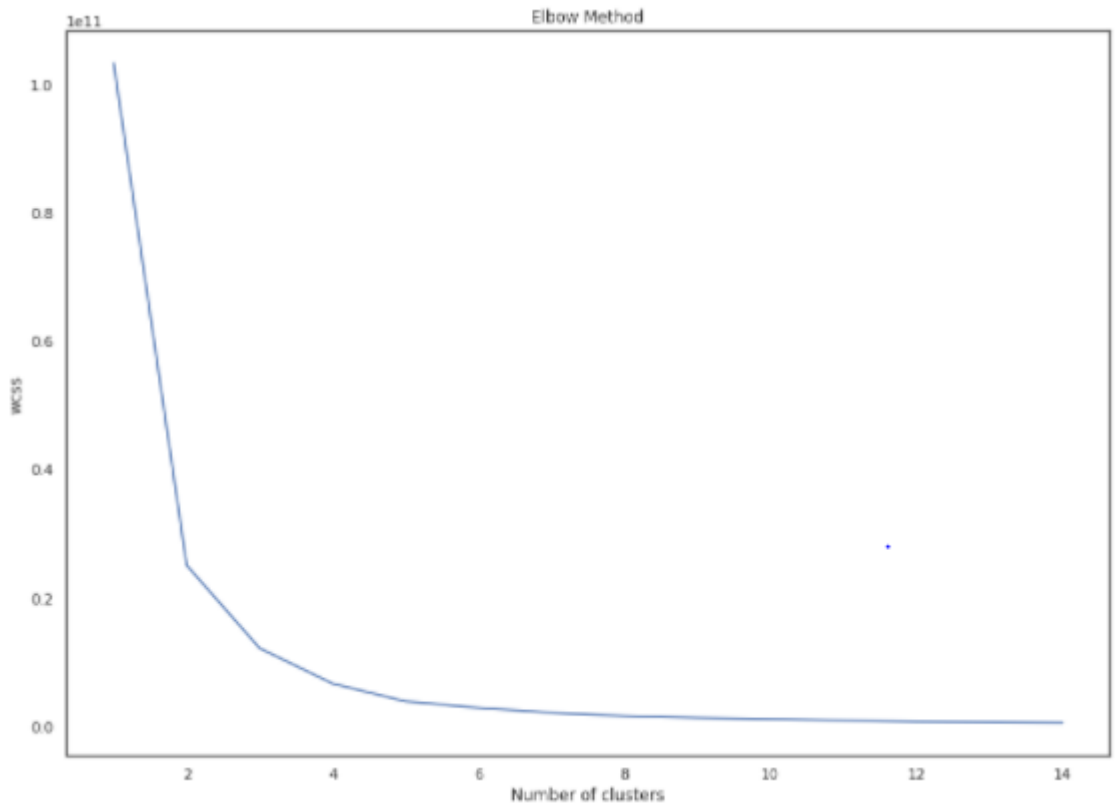
We will use a clustering technique known as *k-means* to do this analysis. The reason for using this technique is that it is appropriate and relevant to our dataset, and it can easily be used in python. K-means is one of unsupervised learning models which means it is used when the data has not yet been labelled or distinguished. (Cui, 2020)

K-means uses a distanced-based clustering to distinguish similar data elements within a cluster from elements in other clusters. All elements are assigned a single unique cluster and cannot be part of two clusters simultaneously. The relevant dataset is divided into K number of groups.

However, we cannot randomly assign or finalize a value for K. We have to ensure that our chosen value for K, minimizes intra-cluster distances while maximising inter-cluster distances. Thereby ensuring homogeneity within a cluster while maintaining the clusters' uniqueness from other clusters.

For this purpose, we use the elbow method by plotting WCCS against the number of clusters. This will aid us in determining the optimal number of clusters we should use for our k-means clustering. WCCS is an abbreviation for Within-Cluster Sum-of-Squares which measures the variance within each cluster. In simpler terms, it measures the distance of the intra-cluster elements to each other, the lower the value the closer each element is to each other within the same cluster. This is the graph that we arrive at:

Figure 1 Graph for Elbow Method



By observing the graph, we can see that the biggest bend can be observed when the number of clusters is 3. To confirm our observation, we use the Kneed package for verification.

```
kl = KneeLocator(range(1, 15), wcss, curve="convex", direction="decreasing")

kl.elbow
```

```
3
```

As we can see in the screenshot above, we have confirmed that 3 is the appropriate number of clusters that we should use for our *k-means* analysis. Consequently, we proceed with executing our k-means model and successfully derive 3 distinct clusters. We assign unique labels to these clusters which are 0, 1 and 2.

We can now view the description of our clusters, although we cannot visualize them. This is because our data has too many dimensions for a two-dimensional visualization to be possible. Visualizations are usually very helpful when interpreting the results of clustering but an analysis in their absence is still, fortunately, possible.

But before we discuss the results, we can internally validate the results of our k-means clustering by looking at the Davies-Bouldin index for our k-means model.

```
from sklearn.metrics import davies_bouldin_score
labels = kmeans_model.labels_
davies_bouldin_score(X, labels)
```

```
0.5141209863841669
```

The Davies-Bouldin index score compares the variation of the data elements within a cluster to the distance between the clusters. Therefore, a lower score indicates that the clustering is relatively good. The score for our clustering comes out to approximately 0.51, which is good. We can now proceed to interpret the results of the k-means clustering.

4.2 Interpretation of K-means Clustering

For the purposes of this report, we will refer to these clusters henceforth as Cluster0, Cluster1 and Cluster2. We can see in the Jupyter notebook that the sizes of the clusters vary significantly. Cluster0 has the majority of the elements with 64731 rows of data. Cluster1 has 15992 rows of data while Cluster2 has 23179 rows of data. This means that Cluster0 is easily the biggest cluster and if we assume that this dataset is accurately representative of the population (the actual number of customers of the airline) then this means they are the primary or major customer segment for the airline.

Let's take a look at the mean values for all three clusters in the next page in the figures below. We are currently looking at the mean values only to get an overview of what the data is representing.

Table 1 Mean values of Cluster0

Age	38.320897
Flight_Distance	539.008373
Inflight_Wifi_Service	2.805075
Departure_Arrival_Time_Convenient	3.290970
Ease_Of_Online_Booking	2.834251
Gate_Location	2.975808
Food_And_Drink	3.152340
Online_Boarding	3.159663
Seat_Comfort	3.291583
Inflight_Entertainment	3.239997
On_Board_Service	3.278250
Leg_Room_Service	3.247106
Baggage_Handling	3.581870
Checkin_Service	3.233726
Inflight_Service	3.587225
Cleanliness	3.199898
Departure_Delay_Min	14.907587
Arrival_Delay_Min	15.439227
Gender_Female	0.510157
Gender_Male	0.489843
Customer_Type_Disloyal Customer	0.244520
Customer_Type_Loyal Customer	0.755480
Type_Of_Travel_Business Travel	0.604795
Type_Of_Travel_Personal Travel	0.395205
Class_Business	0.311875
Class_Eco	0.593302
Class_Eco Plus	0.094823
Satisfaction_neutral_or_dissatisfied	0.670791
Satisfaction_satisfied	0.329209

Table 2 Mean values of Cluster1

Age	42.739557
Flight_Distance	3091.361993
Inflight_Wifi_Service	2.844265
Departure_Arrival_Time_Convenient	3.035063
Ease_Of_Online_Booking	3.006436
Gate_Location	2.996061
Food_And_Drink	3.352124
Online_Boarding	3.794635
Seat_Comfort	3.816221
Inflight_Entertainment	3.684262
On_Board_Service	3.627150
Leg_Room_Service	3.679735
Baggage_Handling	3.765133
Checkin_Service	3.450413
Inflight_Service	3.729217
Cleanliness	3.511148
Departure_Delay_Min	14.570785
Arrival_Delay_Min	14.937822
Gender_Female	0.502439
Gender_Male	0.497561
Customer_Type_Disloyal Customer	0.020135
Customer_Type_Loyal Customer	0.979865
Type_Of_Travel_Business Travel	0.933904
Type_Of_Travel_Personal Travel	0.066096
Class_Business	0.914020
Class_Eco	0.071973
Class_Eco Plus	0.014007
Satisfaction_neutral_or_dissatisfied	0.284955
Satisfaction_satisfied	0.715045

Table 3 Mean values of Cluster2

Age	40.018508
Flight_Distance	1693.593727
Inflight_Wifi_Service	2.816693
Departure_Arrival_Time_Convenient	3.171064
Ease_Of_Online_Booking	2.926692
Gate_Location	2.966650
Food_And_Drink	3.252671
Online_Boarding	3.476846
Seat_Comfort	3.592303
Inflight_Entertainment	3.464950
On_Board_Service	3.504870
Leg_Room_Service	3.482742
Baggage_Handling	3.679710
Checkin_Service	3.400535
Inflight_Service	3.728618
Cleanliness	3.373941
Departure_Delay_Min	14.715993
Arrival_Delay_Min	14.606421
Gender_Female	0.503300
Gender_Male	0.496700
Customer_Type_Disloyal Customer	0.122223
Customer_Type_Loyal Customer	0.877777
Type_Of_Travel_Business Travel	0.758100
Type_Of_Travel_Personal Travel	0.241900
Class_Business	0.641098
Class_Eco	0.310151
Class_Eco Plus	0.048751
Satisfaction_neutral_or_dissatisfied	0.470296
Satisfaction_satisfied	0.529704

We can see that the age is quite similar for all three clusters, it ranges from 38-42 years old, so that doesn't seem to be a distinguishing factor between them. However, flight distance seems to vary significantly over here. Cluster0 has a mean of 539 while Cluster1 has a mean of 3091. Cluster2 is somewhere in the middle with a mean of approximately 1700. This means that the majority of the customers use the airline for short haul flights rather than long haul flights.

Keeping that in mind, there seems to be quite a bit of variation when we compare the satisfaction of the three clusters. The variable for satisfaction which measures the probability of customers satisfied is 'Satisfaction_satisfied'. We can see that for Cluster0 this number is approximately 0.33, which is very low compared to the number of customers which were neutral or unsatisfied which is 0.67. However, this number shoots up to almost 0.72 in Cluster1 while it is 0.53 for Cluster2. A quick deduction can be made here that customers taking short distance flights are significantly dissatisfied while customers taking long distance flights seem to be much more satisfied with the services of the airline. While customers who take flights averaging around 1700 miles are almost split evenly between satisfied and dissatisfied and neutral.

Another thing to notice is that the departure delay and arrival delay in minutes is similar across all three cluster with the number hovering around 14 to 15 minutes. This delay might not feel a lot to customers who are travelling long distances since as a percentage it doesn't delay their flight too much. If someone has a flight for 15-18 hours for example, they will not be too bothered about a 15 minute delay. Also, such delays are somehow expected given the various factors at play for a long haul flight. However, the same 15 minute feels much more of a nuisance when one has only a 75 minutes or 120 minutes flight. One reason could be that most customers want to reach their destination quickly when taking a flight otherwise they could have opted to take alternate forms of transportation such as bus, train, or car. This observation can be further enhanced by seeing that the majority of customers in Cluster0 (60%) are traveling for business purposes. A 15 minute delay for them could mean missing an important meeting or appointment given how tight schedules executives have these days.

Cluster1 though also has customers who are travelling for business purposes (93%) but this could be linked to the fact that 91% of customers in Cluster1 opt for a business class flight. Whereas only 31% of the customers in Cluster0 opt for a business class ticket. This means that an increased level of service cannot be used to make up for the dissatisfaction for the departure or arrival delay of the flight for customers in Cluster0. While customers in Cluster1 do not mind spending extra time in a flight where they are given a an above average service.

We see also that 97% of the customers in Cluster1 are loyal customers of the airline whereas this figure is at only 75% for Cluster0. This could be another reason for the significant disparity of satisfaction figures between the two clusters. Since loyal customers are more likely to forgive small errors while the same can't be said for customers that are not loyal.

4.3 Recommendations based on Clustering results

The airline needs to take measures to address the low satisfaction levels of the major customer group which is Cluster0. As we have seen most people who take these flights travel for shorter distances and usually opt for economy class tickets. This means they have a tight budget and want to reach their destination as soon as possible.

The airline can offer better services to the customers in the economy class by offering a better inflight Wi-Fi service as that is one of service which the customers in this group are least satisfied with an average rating of 2.8. As highlighted earlier, this could be because customers in Cluster0 want to work during their flight and a better inflight Wi-Fi service would make sure they can continue working without stress. Another, thing the airline needs to improve at is to make sure that flight delays are reduced significantly where possible. They could maybe inform the customers in advance if they expect any delays, so that they adjust their expectations accordingly. An automated text and/or email could be sent to the customer as soon as the airline gets to know about any expected delays.

Online Boarding as suggested earlier in this report should be improved significantly to make the customer experience better. Obviously, this will have to be improved for everyone as this improvement cannot be tailored just to the customers in Cluster0.

As we saw that customer of business class flights seem to be very satisfied with the airline's services but customers in Cluster0 might not have the budget to afford the prices. The airline also has an eco-plus class which seems to be severely under used with just above 9% of customers in Cluster0 opting for it. This figure is still much higher compared to Cluster1 with just above 1% of customers opting for it and only 5% of customers in Cluster2. The reason for the low numbers in the latter could be because customers here can afford to buy business class tickets which is why most of these customers prefer that over eco-plus class. The airline could try to reduce prices for the eco-plus class and increase the services offered in this class to entice customers in Cluster0 to upgrade to eco-plus from eco. This would help in ensuring a better satisfaction level in Cluster0.

4.4 Limitations of Clustering Analysis

Our conclusions about Cluster0 taking issue with consistent departure and arrival delays might be incorrect because they do rate Departure_Arrival time convenience (which is 3.2) higher than Cluster1 (which is 3.0) but only marginally. A further survey should be carried out to validate findings reached here. More data could be collected by giving customers more open-ended questions to give them space to voice their opinions. We did not see any such data in the current dataset provided to us.

5. References

Anderl, Eva (2021): *11_Big Data and AI - Winter2021 - Time Series*, Unpublished Script, Munich, Germany: University of Applied Science Munich.

Cui, M. (2020). Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. *Accounting, Auditing and Finance*, *1*(1), 5–8. https://doi.org/10.23977/accaf.2020.010102

Statista (2022): *Share of global mobile website traffic 2015–2021*, Statista, [online] https://www.statista.com/statistics/277125/share-of-website-traffic-coming-from-mobile-devices/ [Accessed on 15.01.2022].

Xu, R., & Wunsch, D. (2008). *Clustering* (1st ed.). John Wiley & Sons.

6. Appendix

Link to the Jupyter Notebook:

https://colab.research.google.com/drive/1FVuYQOmpIS9dm89mnUkUlyJgAY6BgLpz#scrollTo=LlPxgxknjv31