Ein Bild, das Schrift, Text, Grafiken, Symbol enthält.

Automatisch generierte Beschreibung



**Assessment Submission Form**

|  |  |
| --- | --- |
| **Student Number**  (If this is group work, please include the student numbers of all group participants) | GH1021815 |
| **Assessment Title** | Individual Final Project |
| **Module Code** | B141 |
| **Module Title** | Data Mining |
| **Module Tutor** | Prof.Dr. Mahmoudreza Babaei |
| **Date Submitted** | 14.12.2023 |

|  |
| --- |
| **Declaration of Authorship**  I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.  I fully understand that the unacknowledged inclusion of another person’s writings or ideas or works in this work may be considered plagiarism and that, should a formal investigation process confirms the allegation, I would be subject to the penalties associated with plagiarism, as per GISMA Business School, University of Applied Sciences’ regulations for academic misconduct.  Signed………………………………………………. Date ……………………14.12.2023……………………… |

Scope: review, airline name, type of aircraft, traveler type, seat type and ratings of service during the flight, including overall rating. All other columns presented in a dataset I wasn’t using as they will not provide me with additional information for my project.

Objective: Based on all information provided by customer, determine whether he likes the travel

Problem statement: Company for which I’m working provides services with selling avia tickets among all companies. So, based on all reviews from customer, company wants to know whether to recommend this company again to this customer.

In my pipeline I firstly clean all the data, for further processing. This includes also handling with «Nan» values, for text columns there was not so many duplicates, so I just drop them (except type of aircraft, there was many duplicates, and I do not really need this column, so I just drop the whole column). For the numerical columns I made interpolation. Interpolation is a process when a machine is replacing missing value with a number which is between other values (which is presented) presented in this row. I will prove that this method works for this dataset in exploration part.

From the data exploration part, I realized that most of the rows similar to each other. This testifies, that dataset is collected correctly, and we have similar customers which may have some patterns which I would track in next exploration steps. Also, I should mention that ratings are distributed normally, and they all are in range from 0 to 5, so I was thinking about not even scaling them, but scaling increased accuracy, so I leave it, if this pipeline will be used for further investigation, please consider that this step may be deleted.

After the distribution of numerical features, you can see correlation matrix between these features, they are highly and positively correlated, so we could understand that if person didn’t like even one single thing from all flight, he may also didn’t like the other things, even though they were good.

Then you could see correlation matrix with sentiments. Sentiments is a type of review (neutral, positive or negative) predicted by machine. So, based on the words, that were presented in this review machine understands whether it was a positive review neutral or negative. We can see that customers are mostly not satisfied with the flight if they do not have Inflight Entertainment, so for our app we should preferably display companies, which has this engagement.

Then I display companies that are mostly recommended, so they also need to be displayed firstly, and we can also increase our markup for this companies, as well as collaborate with them more.

Then you could see 50 companies with best overall rating, these are also the companies that should be displayed more, and we should care about them, but not as much as for previous 10 companies as there are might be some outliers(such as overall rating is 10 for a user, just because he/she had a good day)  
As my main task was to determine the recommendation of the user, it was a classification problem, so it was solved using 2 classification method(you can read more about the methods used in code part).

The overall accuracy of my project is 96% which is a very good result, and we could really trust that machine could predict the recommendation of our customer. And I also need to mention, that machine is working better in predicting where the customer didn’t like the flight.

As for further recommendation, I would suggest, that when our customer is searching for a flight, we firstly check which companies he might recommend and then display them firstly, with a bit higher markup (for example: markup could depend on overall rating/ difference between the flights with the company which the customer would recommend and which not)

Overall, I could say that this model could possibly highly increase our profit and provides our customers with a comfortable and convenient flight. So this would be win-win strategy, that could be simply implemented and could significantly increase value for our customers and us

