Project Report

MACHINE LEARNING TECHNIQUES FOR PATTERN RECOGNITION APPLICATIONS

Artificial Intelligence (BEJ42803)

Dr. Nor Surayahani

Universiti Tun Hussein Onn Malaysia

08/07/2023

Semester 2 2023

Moritz Hoehnel (JD220004), Mattis Ritter (JD220003)

Table of Contents

1. Introduction

For transport companies the customer satisfaction is the most important issue. A happy customer is not just likelier to return, in times of social media word spreads fast about the quality of the transport [1]. This is why transport companies need to ensure that customers liked their traveling experience. This report focuses on analysing a customer survey. The survey considers different aspects of customer experience like the punctuality. To have more accurate results and consider the feelings of range of different customers, big data sets get analysed. As this is hardly possible with conventional methods pattern recognition will be used [2]. This report presents machine learning techniques to analyse the customer survey with pattern recognition.

The customer survey is about the Shinkansen Bullet Train which is Japan's highspeed train. The objective is to predict if the customer was satisfied with the journey or not. Therefor different prediction models shall be used. It is also important to understand the data and find out which parameter contributes at which weight to the customer satisfaction.

2. Problem Statement

"In 2021, the estimated number of domestic travellers using Shinkansen high-speed railways in Japan totalled around 34.52 million" [3]. To find out if they had a satisfactory trip random people were asked to perform a survey. The results of this survey need to be analysed. A total of 94379 customer data with 13 categories each need to be looked at. The issue is to find a prediction to know if a customer is satisfied with the overall experience.

3. Methodology

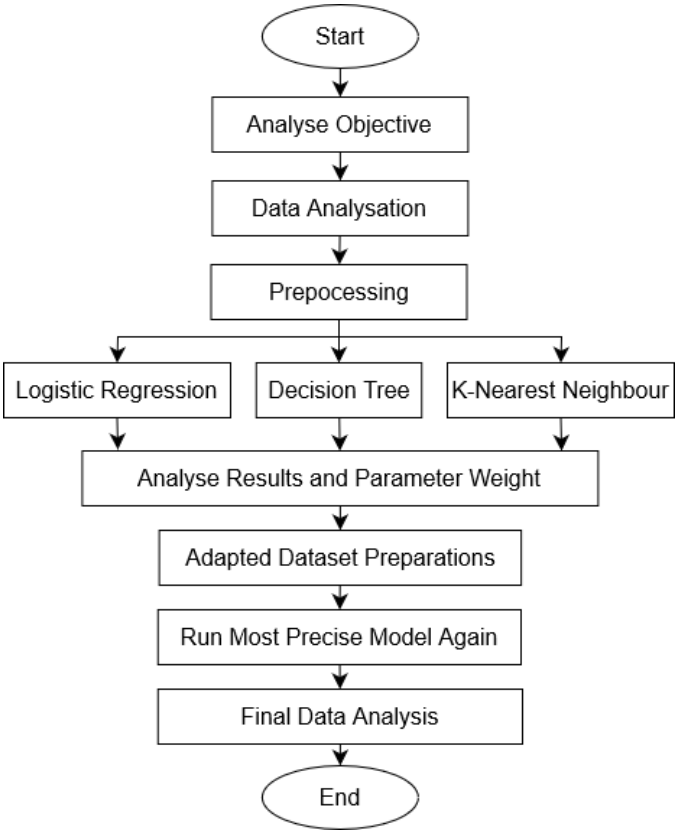

Figure 3.1: Methodology Flow Chart

Figure 3.1 gives an overview about the methodology of the project. The flow chart shows that after the project start the objective got analysed. For the following steps the project team used the Jupyter Notebook. After the dataset was loaded into the notebook, the set was analysed concerning its structure and content. For more details on the dataset see chapter 4.

This helped to prepare the dataset during preprocessing. All rows with null values were removed as they contain missing data and the models are not able to work with that. A major disadvantage of this method is that the dataset gets smaller. Next, unsignificant parameters were detected and removed. For this case it is only the ID. In order to encode the categorical object variables, the unique values of each row were detected and then ranked from extremely poor to excellent, by assigning the numbers 0 to 5 to them. The exact mapping is the following:

'Extremely Poor': 0, 'Poor': 1, 'Needs Improvement': 2, 'Acceptable': 3, 'Good': 4, 'Excellent': 5

For gender male got encoded as 0 and female as 1 and for the seat class green car got encoded as 0 and ordinary as 1. Now the dataset is split into input x and output y and then also into 70% training and 30% testing data. To ensure that there is now bias in the train or test set, the percentage of classes got normalized.

Afterwards the flow chart divides into three branches. This is where the first round of model running was performed and each model's performance got checked. As far as the notebook is concerned, the models are fitted in the following order: Logistic Regressions, Decision Trees and K-Nearest Neighbour. As they rely on the same data set, they can be performed in any order. To find the best prediction for the data, the Logistic Regression was run with unscaled and scaled data, the Decision Tree with three different settings (default, weighted and with entropy and max depth 15) and the K-Nearest-Neighbour was run once with scaled data and the number of neighbours being 9.

In the step of 'Analyse Results and Parameter Weight' the first phase was to scrutinise the output of the test set with focus on accuracy as described further in chapter 5. Based on the results the most precise model was chosen, which was the decision tree with entropy and max depth 15. The model was reused and fitted to differently prepared datasets. Firstly, two new approaches of data encoding were used, one by adding dummy values for all parameters another by encode null values with "-1". In a last adaption of the dataset the two parameters with the lowest importance got dropped and the rest of the encoding was handled as in the beginning. After that a final data analysis was performed. Here one can also find the comparison of the different model runs.

4. Dataset Descriptions

The following chapter describes the dataset. To understand the dataset the functions ".head()", ".tail()", ".info()" and ".describe()", had been applied as well as it was inspected for null values. The ".tail()" function reveals that the dataset has 94379 entries. The ".info()" function shows that the dataset has 14 columns. The first one is the survey "ID" and the last one is the "Overall_Experience" which evaluates if the passenger liked the journey. The customer can either like or dislike, which is expressed by a "1" or "0" respectively. To analyse the remaining categories figure 4.1 shows a summary of the most important dataset information.

#	Column	Dtype	Null Count
0	ID	int64	0
1	Departure_Delay_in_Mins	int64	0
2	Arrival_Delay_in_Mins	int64	0
3	Gender	object	77
4	Seat_Comfort	object	61
5	Seat_Class	object	0
6	Arrival_Time_Convenient	object	8930
7	Onboard_Wifi_Service	object	30
8	Ease_of_Online_Booking	object	73
9	Baggage_Handling	object	142
10	Legroom	object	90
11	CheckIn_Service	object	77
12	Cleanliness	object	6
13	Overall_Experience	int64	0

Figure 4.1: Summary of Dataset Description

Under "#" each column has a number aligned. The column "column" shows the names of the categories. "Dtype" is short for datatype. One can see that the datatype is either and integer or an object. For creating a model, all objects need to be changed to numbers, which is happening during preprocessing. "Null Count" is displaying the number of rows for each column that do not have a result inserted. Rows that have one or several empty fields also need special handling in the preprocessing. as already described in chapter 3.

5. Results

Every model is evaluated with a metric function that prints the classification report and displays a confusion matrix for further visualization of the results. This chapter will only focus on the best performing model, that is determined as described in the following paragraph.

Table 5.1: Comparison of all Models

Nr.	Model Description	Data Preprocessing	Accuracy
9	Decision Tree with Entropy and max Depth 15	Drop Rows with Null Values Drop least important Features Encode Categorical to Numerical	0.884
5	Decision Tree with Entropy and max Depth 15	Drop Rows with Null Values Encode Categorical to Numerical	0.881
8	Decision Tree with Entropy and max Depth 15	Drop Rows with Null Values Create Dummies for Categorical	0.879
7	Decision Tree with Entropy and max Depth 15	Replace Null Values with '-1' Encode Categorical to Numerical	0.878
6	K-Nearest-Neighbour	Drop Rows with Null Values Encode Categorical to Numerical	0.862
3	Decision Tree	Drop Rows with Null Values Encode Categorical to Numerical	0.858
4	Weighted Decision Tree	Drop Rows with Null Values Encode Categorical to Numerical	0.858
1	Unscaled Linear Regression	Drop Rows with Null Values Encode Categorical to Numerical	0.759
2	Scaled Linear Regression	Drop Rows with Null Values Encode Categorical to Numerical	0.759

The most important metric for the model performance is the accuracy as it represents the ratio of the sum of true positives and true negatives out of all predictions. A higher accuracy indicates a better performance. When there is a high difference between precision and recall it is recommended to chose the weighted average of the f1-score for model comparison, but this is not the case here. In the table 5.1 all tested models are ranked based on their accuracy score.

Model 9 has achieved the highest accuracy. Therefore the results of this model will be explained in more detail. The first 3 levels of the model are depicted in figure 5.1.

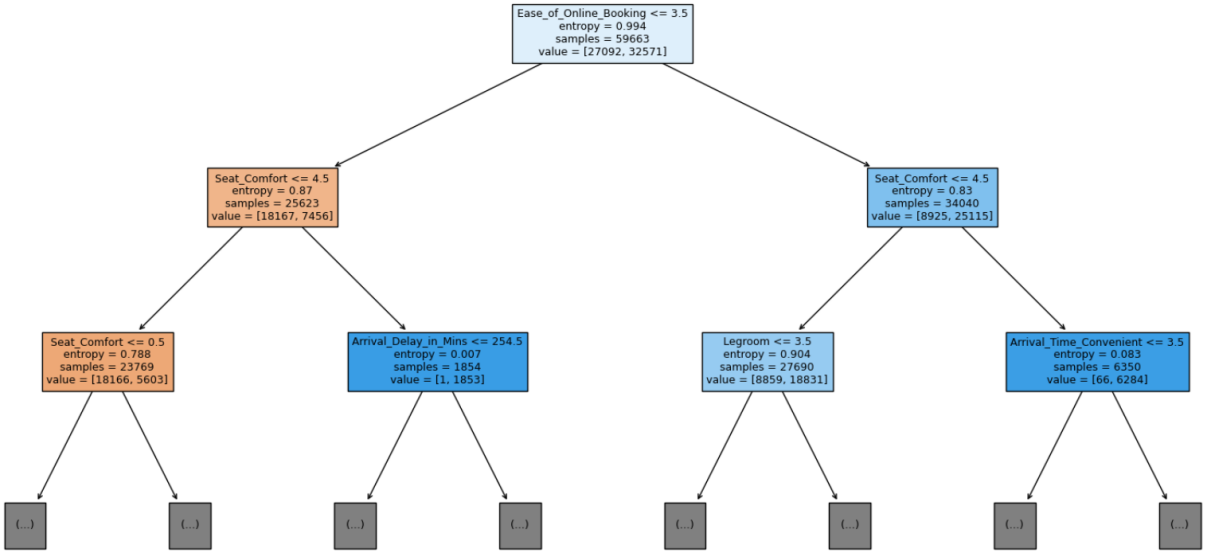

Figure 5.1: Model 9

There are four possible outcomes of a prediction, that are used to calculate the metrics:

1. True Negative (TN): the customer was unsatisfied (0) and the model predicted a bad overall experience (0)
2. True Positive (TP): the customer was satisfied (1) and the model predicted a good overall experience (1)
3. False Negative (FN): the customer was satisfied (1), but the model predicted a bad overall experience (0)
4. False Positive (FP): the customer was unsatisfied (0) and the model predicted a good overall experience (1)

When checking the performance with the training dataset all metrics are approximately 92% which shows that the model is not overfitted to the training dataset, but still quite accurate. This is proven when reviewing the classification report for the test dataset, shown in figure 5.2, as all metrics are still around 88%.

	precision	recall	f1-score	support
0	0.86	0.88	0.87	11622
1	0.90	0.88	0.89	13969
accuracy			0.88	25591
macro avg	0.88	0.88	0.88	25591
weighted avg	0.88	0.88	0.88	25591

Figure 5.2: Classification Report for Model 9

As the precision score shows, 90% of the predictions for a good overall experience where correct and the predictions that the customer was unsatisfied were correct in 86% of the cases. For class 1 precision calculated with the following formula:

$$Precision = TP/(TP + FP)$$

The recall is 88% for both classes, meaning that 88% of the actual cases were caught by the model. For class 1 the formula is:

$$Recall = TP/(TP + FN)$$

The f1-score is a weighted harmonic mean of precision and recall and for this model it is 87% for bad and 89% for good overall experience, meaning that the model is slightly better in detecting satisfied customers. The calculation for the f1-score follows the following formula:

$$F1\ Score\ =\ 2 \cdot (Recall \cdot\ Precision)\ /\ (Recall\ +\ Precision)$$

Accuracy is at 88% and calculated by the following formula:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

For each of these metrics there is also a macro average and a weighted average of the classes displayed, being 88% each. The column support shows how many samples exist for each class.

The confusion matrix, illustrated in figure 5.3, visualizes the absolute number of true negatives/positives and false negatives/positives. As these are used for the calculation of the metrics it is a easy overview of the model's performance.

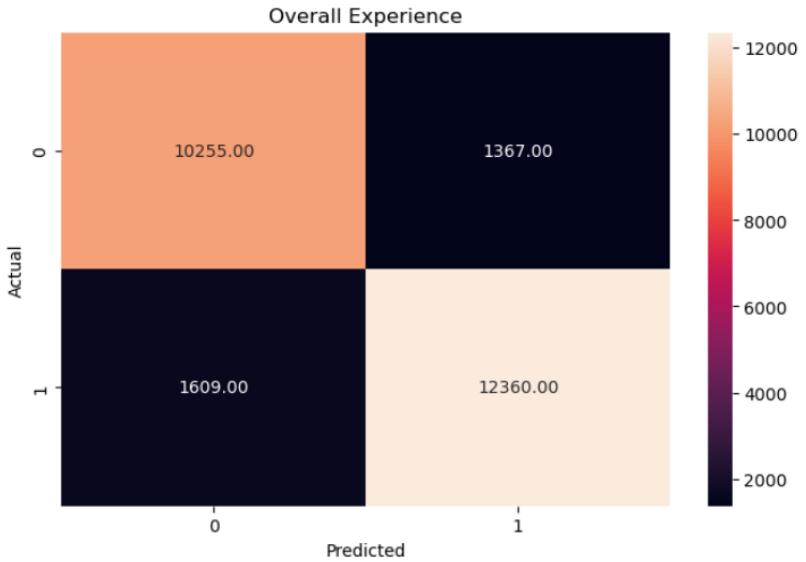

Figure 5.3: Confusion Matrix for Model 9

6. Analysis and Discussion

This chapter discusses the meaning of the model results. As Model 9, the decision tree with criterion entropy and maximum depth 15, scored the highest accuracy, it is chosen as the model of usage. The analysis will focus solely on model 9.

For transport companies it is important to know which categories are the most important for their customers. The Shinkansen team can focus to improve the experience for certain aspects. This will lead to more customers being satisfied. To extract this information a Feature Importances bar graph was created, as shown in figure 6.1.

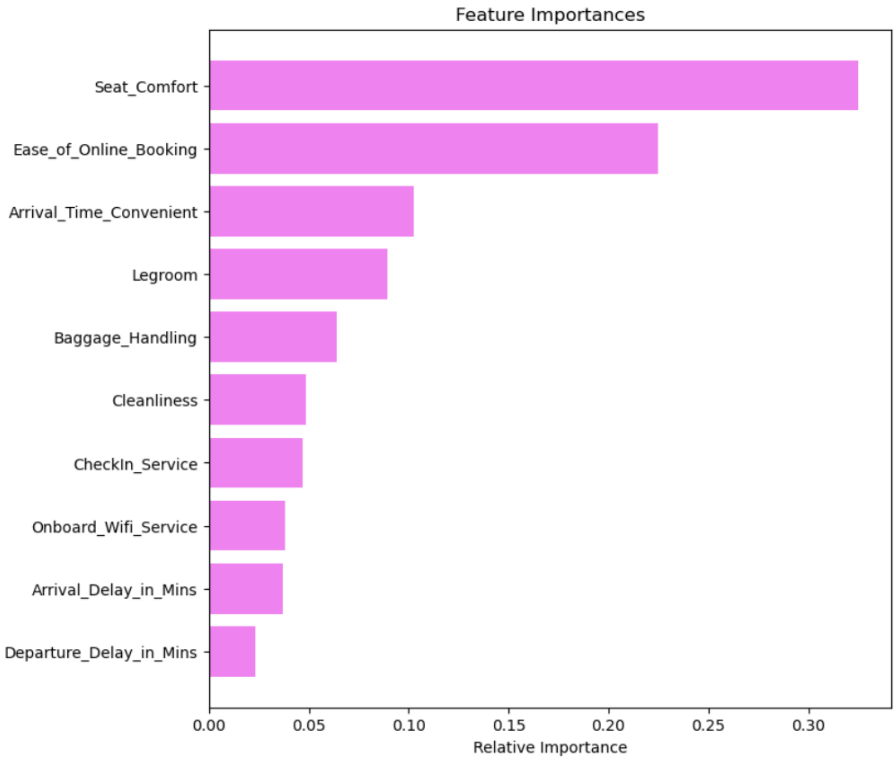

Figure 6.1 Feature Importances of Decision Tree

The bars displayed in figure 6.1 indicate the Relative Importance of each of the categories. The top five categories include “Seat_Comfort”, “Ease_of_Online_Booking”, “Arrival_Time_Convenient”, “Legroom” and “Baggage_Handling”. It is also visible that the first two categories cumulated make up over 50% of the Relative Importance. This means that these two aspects need special attention by the transport company. It is also interesting to see that the two categories for delay have the lowest importance, despite Japanese are known for taking punctuality seriously. This can rule from the fact that at least 50% of the trains do not have a delay and if there is a delay it is only little. Only some trains have huge delays. The delay information have been acquired by using the “.describe()” function.

	Departure_Delay_in_Mins	Arrival_Delay_in_Mins
count	94379.000000	94379.000000
mean	14.638246	14.948463
std	38.128961	38.377695
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	12.000000	13.000000
max	1592.000000	1584.000000

Figure 6.2: Delay Information

As business recommendations the team would encourage the transport provider to keep special focus on their seats and their online booking system. This means for new trains to spend more money on comfortable seats that have as example a convenient head rest. But also service the seats regularly. As the online booking is also important it is advisable to have an own tool that is easy to use, but also open the selling of tickets to other booking platforms a customer may be already used to.

A further advice is to keep the punctuality at the current level to not risk a decline in user experience. At the moment it is not a issue, but if trains get more delayed it is expected that the customers become disappointed.

7. Conclusion

The project had two objectives, the first one was to find the important categories that contribute to a good overall experience. The second one was to predict if a customer will be satisfied. The first objective was achieved by analysing the Feature Importances graph that showed that seat comfort and online booking experience are the two most significant parameters. The second objective was achieved by testing different models and analysing their accuracy to finally use the best model to predict if a customer is satisfied. As both goals are reached the project is a success.

The project team is satisfied with the model accuracy and would recommend using it. Furthermore, the team would like to point out that the two most important categories should be looked at more closely in future.

8. References

[1] S. Kumar and M. Zymbler, "A machine learning approach to analyze customer satisfaction from airline tweets", *springeropen.com*, Jul. 17, 2019. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0224-1 [Accessed: Jun. 27, 2023]

[2] R. Szeliski, *Computer Vision*. Seattle: Springer Cham, 2022.

[3] A. Arba, "Number of domestic Shinkansen travelers in Japan 2016-2021", *statista.com*, Oct. 15, 2022. [Online]. Available: https://www.statista.com/statistics/1272144/japan-tourism-domestic-shinkansen-traveler-number/ [Accessed: Jun. 27, 2023]

[4] Pandas, "User Guide", *pydata.org*, 2023. [Online]. Available: https://pandas.pydata.org/docs/user_guide/index.html [Accessed: Jun, 29, 2023]

[5] scikit-learn, "User Guide", *scikit-learn.org*, 2023. [Online]. Available: https://scikit-learn.org/stable/user_guide.html [Accessed: Jun, 29, 2023]