Stakeholder Report

Introduction / Problem Statement

This assignment is based on a dataset from AirBnB in Singapore, the dataset was updated on 28 August 2019. There are 7907 samples. In this stakeholder report, we will walk you through the most essential analyses and results from python, so you can get an understanding of our findings.

We have made a problem statement, which will guide the analyses of our assignment. The problem statement is*:*

> *To what extent does the accommodation features and titles have influence on the prices?*

To answer the above problem statement we will use Machine Learning techniques on six different features, which might have an influence on the prices to see how precise we can predict whether the accommodation has a prices above or below the average price. We will also use Deep Learning techniques on the accommodation titles to see how precise these predictions are compared to the actual prices and the first predictions.

Based on our prior knowledge, we were expecting that our deep learning model will be more accurate than the machine learning model. The reason for this assumption is that we believe that the title of the accommodations as a feature in the dataset contains more relevant information for the traveller than for example the availability and number of reviews, given that this can vary in many different ways and often time, the reviews are not present in the add. Yet, it is interesting to figure out whether or not this assumption is true.

Preprocessing / the data

The following variables is the foundation of the assignment and the predictions.

Target variable:	Feature variables for Baseline Model	Feature variable for Deep Learning Model
• Price	• latitude • longitude • number_of_reviews • minimum_nights • availability_365 • calculated_host_listings_count	• Name

To make it possible for us to compare the data in the different columns, we needed to scale our data. The idea behind StandardScaler is, that it will transform our data such that its distribution will have a mean value 0 and standard deviation of 1. By doing this on our feature variables, we made it possible for us to compare for example Longitude/lattitude to minimum_nights. By doing this we also prepared the data for for the baseline model.

Variable	Price	Latitude	Longitude	number_of_reviews	minimum_nights	availability_365	calculated_host_listings_count
count	7907	7907	7907	7907	7907	7907	7907
mean	169	0.00	0.00	0.00	0.00	0.00	0.00
std	340	1.00	1.00	1.00	1.00	1.00	1.00
min	0	-2.3	-4.6	-0.43	-0.39	-1.43	-0.61
25%	65	-0.6	-0.30	-0.43	-0.39	-1.06	-0.59
50%	124	-0,1	0.01	-0.36	-0.34	0.35	-0.46
75%	199	0.3	0.54	0.09	-0.18	1.00	0.13
max	10000	4.6	2.85	10.4	23.3	1.06	3.58

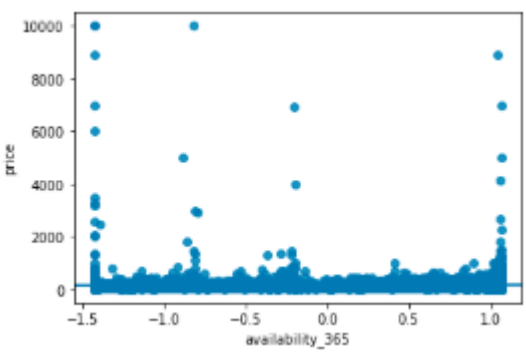
The table above shows our chosen variables where the feature variables has been scaled. The table also shows the mean of the different variables. The mean price of the accommodations is 169 Singapore Dollars.

Linear regression

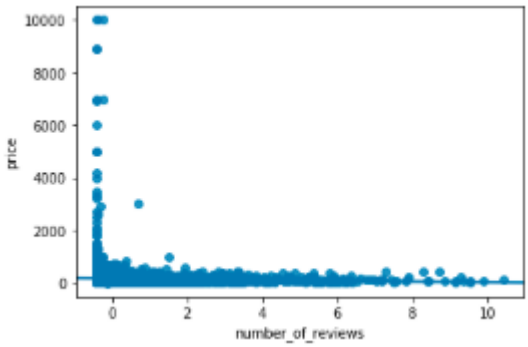
This assignment started with creating some linear regression with the idea of getting an overview of the link between the target value, in this case 'price' and the different features that have been chosen. Since these Linear Regressions were made to give a quick insight at the link between the variables, it will also be possible to get the first insight into what could be the answer to our problem statement.

In the following, we will only visualize three of the linear regressions, due to the maximum amount of pages in this report.

Price and availability



Price and number of reviews



In the above linear regressions, we can interpret that the features does not have an big impact on the price of the accommodation. As an example the regression with the number of reviews has not a big impact on the price. As we can see in the plot, there are a lot of expensive accommodations with no reviews. There could be some potential outliers here, it seems weird that some of the accommodations costs 10,000 Singapore dollars, which is approximately 50,000 DKK.

Supervised Machine Learning

Supervised learning is where you have input variables and an output variable and you use an algorithm to learn the mapping function from the input to the output. The goal is to approximate the mapping function so well that when you have new input data that you can predict the output variables for that data.

To be able to answer our problem statement, we wanted to predict whether the prices of the accommodation was below or above average. As mentioned we used six different features

The result show that our accuracy is approximately 75 percent, which is a decent result, but one could have hoped for a bit more. However, this do makes sense, as our EDA showed that there were little link between our chosen features and the price (below or above average) that people pay for renting a apartment/room in Singapore.

Furthermore, one can argue that it can be difficult to predict, since many features such as size, access to bath and kitchen, location and room type could be relevant with regards to the price, which this model cannot capture. Despite the fine result, we hope that our predictions within the deep learning part will give us a higher accuracy.

Deep Learning

In this assignment, the overall architecture is Convolutional Neural Network (CNN). Then the LSTM, CuDNNLSTM and GRU models have been used to investigate the problem statement. The reason for using three different models are to get an overview of, which model predicts the best result.

We started with using the LSTM model, which gave us 0.39 loss. To tune this model, we tried to change the epochs to 10 and 100 and with a batchsize of 5 and 200, which increased the loss of the function and accuracy was more or less unchanged, which was not the desired result. Therefore, we used the CuDNNLSTM to see, if we could get a better result with this model. By doing this, we got a loss of 0.57, which is far worse than the LSTM model, which is why this model is not desired to use. Therefore, did our tuning of the models not improve our results.

Model Type	Loss	Accuracy
LSTM	~ 0.38 - 0.42	~ 78.34% - 84.89%
CuDNNLSTM	~ 0.52 - 0.58	~ 80,43% - 83.82%
GRU	~ 0.53 - 0.59	~ 74.9% - 84.95%

All three predictions models within the deep learning part of this assignment had a high accuracy as result. The reason for this could be that the chosen feature in this part, which was name of the accommodation, gives a great description of the details of the accommodations, which is beneficial for the models to work with. To give an example "*2-bedroom luxury penthouse + Jacuzzi + BBQ deck*" has a price above the average mean price, whereas "*Small room for you"* is below the average mean price.

Above results show that our predictions within the deep learning part of the assignment are quite good given that we got an accuracy of above ~ 80 percent.

Conclusion

This assignment shows that our deep learning predictions are higher than the supervised Machine Learning prediction (~ 83% vs. 75%). This might be because the Machine Learning model is based on features, which do not have a very high impact on the price of the accommodation. In the Deep Learning part, there has been used three different models to predict on AirBnB prices according to the title of the accommodation. The reason for choosing three different models to do the same predictions is that we wanted to investigate, which could deliver the best results by comparing the loss and accuracy of each model. The results show that our LSTM model predicted with the best accuracy and lowest loss score. However, the numbers were quite close and similar.

We believe that the reason that our Deep Learning models make better predictions is that the input in these models are based on the name which might include some more relevant information about the accomodation. For example location, number of rooms and description of the whether it is a luxury or studio accommodation, which were shown in previous examples.

Therefore, it can be concluded that with this dataset and our problem statement, the Deep Learning models are the best choice, since they have the highest accuracy.