FACULTY OF SOCIAL SCIENCES
Department of Economics
University of Copenhagen

Social Data Science – Summer School 2018

Mads Berner Bruun (RVJ233), Jonathan Lasse Andersen (QWK638),

Sofie Juel (GDH466) and Anna Beck Thelin (JVS499)

Predicting Hotel Room Prices in New York City

ECTS points: 7.5

Date of submission: 01/09/2018

Keystrokes: 26.664

Assignment contribution

Introduction

Line 1-6: JVS499

Line 7-14: GDH466

Line 15-23: QWK638

Line 23-33: RVJ233

Theory

Line 1-15: JVS499

Line 17-36: GDH466

Line 37-57: QWK638

Line 58-78: RVJ233

Data

Line 1-15: JVS499

Line 16-30: GDH466

Line 31-46: QWK638

Line 47-58: RVJ233

Descriptive Analysis

Line 1-7: JVS499

Line 8-14: GDH466

Line 15-21: QWK638

Line 22-29: RVJ233

Empirical Analysis

Line 1-13: JVS499

Line 14-26: GDH466

Line 27-40: QWK638

Line 41-54: RVJ233

Discussion

Line 1-11: JVS499

Line 12-23: GDH466

Line 24-35: QWK638

Line 36-46: RVJ233

Conclusion

Line 1-3: JVS499

Line 4-6: GDH466

Line 7-9: QWK638

Line 10-13: RVJ233

Contents

1 Introduction

New York City is among the cities with the largest hotel markets in the United States, only surpassed by Orlando and Las Vegas. With NYC experiencing a massive boom in 2016, welcoming almost 61 million tourists, more than 5000 new hotels opened in 2017 (BJH Advisors (2017)). Not only is the size of the hotel market in NYC impressive, the hotel occupation rate is around 86 percent, which is significantly higher than the national average (Rothstein (2018)).

With hotels spread across the entire NYC, including Manhattan, Staten Island, Bronx, Brooklyn and Queens, the hotels are widely diverse and attract all types of tourists with different preferences. The most important aspect of hotel marketing strategy is the price setting since this can be done fairy flexible. It is also one of the top determinants for a potential customer's decision to book (Lockyer (2005)). According to M. Collins (2006), the primary influencing factors for the pricing decision are star rating, management type, location, size and amenities..

This paper seeks to identify the optimal regression model to predict the price of a hotel room in New York City and further investigate which features that determine the price. This is done by scraping Expedia for all data on hotel rooms, within the entire price-specter, suitable for two people. The features in scope are star-rating, location and amenities which will be the top 7 amenities that Statista (2017) found U.S. travelers are willing to pay extra for. These are: free WiFi, room service, wellness and spa, pet policy, laundry/ironing service, premium TV and car park. Furthermore, the analysis also includes reviews on Expedia (number and score), free breakfast and free cancellation since these are also expected to have a significant effect on prices. Using machine learning regression models, we investigate which effect different amenities, hotel location and hotel star-rating have on the price of a room. From the empirical analysis, it was found that the Random Forest regression model was the preferred model to predict hotel room prices in NYC, though the model seemed to be overfitting. Furthermore, it was found that the order of the most important features for price setting is star-rating, location, Expedia reviews and lastly amenities.

The paper proceeds as follows: Section 2 goes through the theory behind the different regression models used in the analysis. Section 3 presents the data used in the analysis, section 4 is the descriptive analysis and Section 5 the empirical analysis. Section 6 discusses the results and section 7 concludes.

2 Theory

The purpose of this section is to outline the theory behind the models used to predict hotel room prices. The regression models used in this analysis are OLS, Ridge, Lasso and Random Forest. To enable comparison between the linear OLS regression model and machine learning models, root mean squared error (RMSE) is used as model performance benchmark (Raschka (2015)).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

RMSE, which denotes the difference between predicted, $\hat{y}_i$, and observed values, y_i, ensures fair comparison of different regression models - lower RMSE means better prediction of hotel room prices.

2.1 OLS regression model

The ordinary least squares (OLS) regression model minimizes the sum of the squared residuals. Under the Gauss Markov assumptions (Wooldridge (2012)), OLS is the best linear unbiased estimator. The following expression is minimized:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 = RSS \tag{2}$$

The expression summarizes and squares the difference between the observed and the predicted value of the independent variable.

2.2 Machine learning regression models

This section provides a brief introduction to machine learning. Furthermore, it presents the specific machine learning models used in this paper.

The Hold Out Method

Machine learning trains a statistical model with subsets of the total data sample. The total data sample is randomly split into three subsets of training-, validation- and test

data, each containing 1/3 of the total sample. This will improve the predictive power as subsets are validating the generalized performance of the model.

Hyperparameters

The different learning algorithms provided by the scikit-learn python library come with some default parameters. However, these may not be the optimal parameters of choice for the specific problem at hand. Hyperparameters are parameters that can be used specifically to tune the model. Once the *optimal* values of the hyperparameters are selected, the model's generalized performance is tested on the test-subset of the data. In this analysis, a hyperparameter, λ, has been defined for the Lasso- and Ridge regression models. The machine learning models are trained given each value of λ, and a root means squared error for each is returned. The λ that returns the lowest RMSE is the optimal value.

K-Fold

In order to further increase the performance of the models, the K-fold cross validation technique is applied. This technique splits the training data into k-number of bins, where $k-1$ bins are used for training and the last bin is used for testing. The machine learning models are trained for each of the bins for all the hyperparameters. Hence, for each hyperparameter there are now k-numbers of RMSEs. Taking an average of the RMSEs related to each hyperparameter, the optimal hyperparameter is the one that returns the lowest mean-RMSE.

2.2.1 Ridge regression model

The Ridge regression model is a penalty function which works by reducing the size of coefficients. In that sense, the Ridge regression model addresses the issue of multicollinearity between the features in a model by minimizing the estimates of highly correlated feature variables. With ridge regression the following expression is minimized:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=i}^{p}\beta_j^2 = RSS + \lambda\sum_{j=i}^{p}\beta_j^2 \tag{3}$$

Where p is the number of explanatory variables and n number of estimators. The first part of the expression is the residual sum of squares (RSS), which is what is minimized in an OLS regression. The Ridge regression model then further restricts the parameters β_j by introducing the penalty term λ, which is a hyperparameter. Hence, if $\lambda = 0$ the Ridge regression model yields the same result as the OLS. If β_j is unconstrained, they may explode and are therefore susceptible to very high variance. λ is then used to regularize and control the size of the coefficients. The K-fold technique is applied when training the Ridge regression model. The λ related to the lowest mean-RMSE is the optimal penalty value.

2.2.2 Lasso regression model

An alternative to the Ridge regression model is the Lasso regression model. The Lasso regression model differs from the Ridge regression model by not only punishing high values, but actually setting them to 0 if they are irrelevant. Hence, you might end up with fewer features after carrying out the regression than included in the initial model, which can be a great advantage. Lasso minimizes the following expression:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=i}^{p}|\beta_j| = RSS + \lambda \sum_{j=i}^{p}|\beta_j| \tag{4}$$

The only difference between the minimization problem in Ridge (3) and Lasso (4), is that the regularization term β_j is in absolute value. Here the K-fold technique is also applied, and the λ related to the lowest mean-RMSE is likewise the optimal penalty value.

2.2.3 Random Forest regression model

Before the Random Forest regression model is introduced, it is necessary to outline the basics of Decision Tree regression.

A decision tree can be used both for classification- and regression models. Each tree is build up by the same procedure in that it has a root node which is the best predicting feature. From this node, the tree splits up into a series of child notes. Each internal node of the tree is a feature from the dataset and the tree will end with a node called leaf nodes. That way, a decision tree is build by iteratively splitting up the internal nodes until all the leaf nodes observations have been allocated, or until a specific criterion has been satisfied.

The Decision Tree regression seeks to minimize the mean squared errors (MSE) of each node. The MSE of the Decision Tree regression can be described as:

$$I(t) = MSE(t) = \frac{1}{N_t} \sum_{i \in D_t} (y_i - \hat{y}_t)^2 \tag{5}$$

Here, N is the number of training samples at each node t, D is the training subset at node t, y_i is the true value from data and $\hat{y}$ is the predicted target value sample mean, that is defined as:

$$\hat{y}_t = \frac{1}{N} \sum_{i \in D_t} y_i \tag{6}$$

The Random Forest regression model differs from Decision Tree regression model by having multiple decision trees. This typically means that the predictive performance of a Random Forest regression will exceed the predictive performance of an individual decision tree since the Random Forest regression is able to capture more of the model variance and make better estimators. To enable comparison, the MSEs are converted into RMSEs. The number of features related to the lowest RMSE is the optimal number of features.

3 Data

This section outlines how data on hotel prices in New York City is gathered by scraping Expedia.com and how this data is then prepared for analysis.

3.1 Data collection

The data used in this analysis is scraped from Expedia. The data holds information about hotel rooms in New York City suitable for two people during week 19, 2019. This week is assumed to be representative for an average week in NYC. The data is scraped once for each day, leaving us with 7 datasets that are merged into one large dataset. The search results on Expedia exclude prices on hotel rooms that are sold out. By scraping once a day, there are as many of the hotels included in the analysis as possible during the given week. Also, this makes it possible to investigate the effect of weekend days (Friday to Sunday) on the price-level.

To automate the web scraping processes and to get as much data about each hotel room as possible, a script that uses two different scraping techniques in conjunction is used. First, the script accesses Expedia's hidden API through the browser's network monitor and then gains access to each of the search-hits which contain all data on each hotel room. This data is collected by extracting the URLs for each of the search-hits and parsing these into a Python library [1] specifically for working with HTML. Using the library, details on hotel and room amenities, star-rating etc are extracted. The raw data now contains information on room- and hotel amenities, star-rating, guest reviews (amount of reviews and review score) and hotel longitude/latitude that will be used to determine the geographical location of the hotel.

[1] The Python library used is BeautifulSoup.

3.2 Data cleaning

After the Expedia data has been scraped, duplicates and outliers are excluded from the dataset. First, Expedia displays sponsored hotel rooms several times in one search. This means that some hotel rooms are repeated up to 30 times per day. Secondly, a small number of hotels list the same room with two different prices on the same day. Since the purpose of this paper is to predict the price of the best-fit hotel rooms, the most expensive instance of these reoccurring rooms are removed from the dataset.
The chosen amenities are converted into dummy variables which take on the value 1 if the amenity is provided at the given hotel and 0 if it is not. Furthermore, it is assumed that if a value is missing, the amenity is not provided and the value 0 is returned.
Also, looking into the distribution of the hotel prices revealed one major outlier of 8999 USD per night, which was removed from the analysis.

Lastly, rows with missing values in any of the explanatory variables are dropped. We then end up with a dataset consisting of 2838 rows, containing information on 490 hotels.

3.3 Dependent variable

Hotel room price is the dependent variable of this paper. Expedia's search engine often displays two prices for each room; a normal price and a discount price. Since discount prices are determined by a range of factors unrelated to hotel features and amenities (such as special Expedia offers) this paper will only be considering the original Expedia price of the hotel rooms. In order to obtain a normal distribution of the dependent variable, the price variable is log transformed which leads to a better overall model fit.

3.4 Explanatory variables

The explanatory variable are chosen based on evidence regarding hotels' price decision determinants, along with an investigation on what hotel amenities U.S. travelers are willing to pay extra for. An overview of the variables in the final dataset is found in Appendix table (4). The hotels are assumed to primarily set their prices according to their star-rating, location and amenities. Star-rating is given based on criteria formulated by Expedia.com[2]. The star rating of the hotel increases with the amount and quality of hotel

[2] https://www.expedia.com/Hotel-Star-Rating-Information

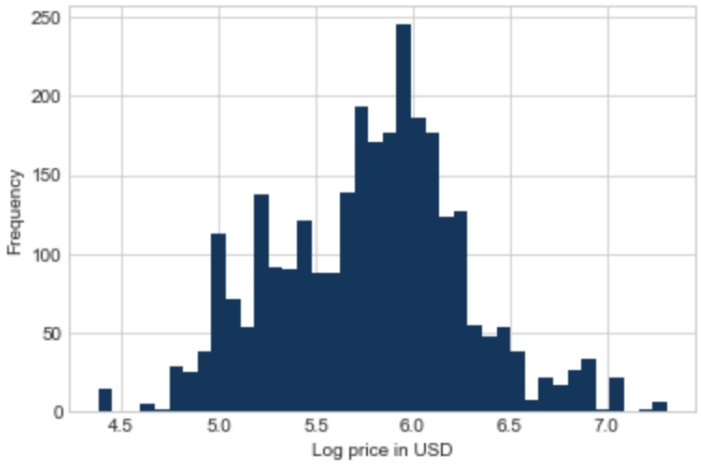
amenities, why it could be expected that these were correlated. However, the amenities used in this analysis are offered by hotels in the entire star-rating scale. Furthermore an investigation of the correlation matrix and Eigen values of the explanatory variables show no signs of multicollinearity.

With regard to hotel location, the analysis will be considering each hotel's distance to Times Square as well as which city borough each hotel is located within. However, since this information could not be scraped directly from Expedia, python packages for geo-spatial analysis were used to generate the variables. First the hotel coordinates (longitude/latitude data) were used to find the distance to Times Square. Second, the coordinates of each hotel were looked up in a shapefile with polygons that define the geographical boundaries for each of the five boroughs in New York City. The newly generated variable for city boroughs is further split into five dummy variables: Manhattan, Brooklyn, Queens, Bronx, Staten Island.

4 Descriptive Analysis

This section presents the central descriptive statistics on the cleaned data. The final dataset contains information about 2838 hotel rooms across New York City's five boroughs.

Figure 1: Distribution of log transformed prices



The histogram in figure 1 shows the distribution of the log transformed prices. With a mean value of 5.78 and a median value of 5.83, the distribution is slightly skewed to the left. The normal distribution around the mean supports the decision to work with log transformed prices in the predictive models. Across the entire sample, the untransformed mean price is 365 USD per night.

Figure 2: Relationship between price and number of stars

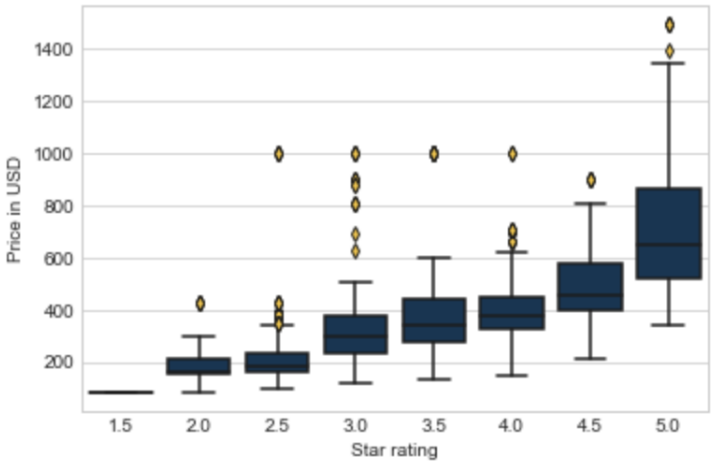
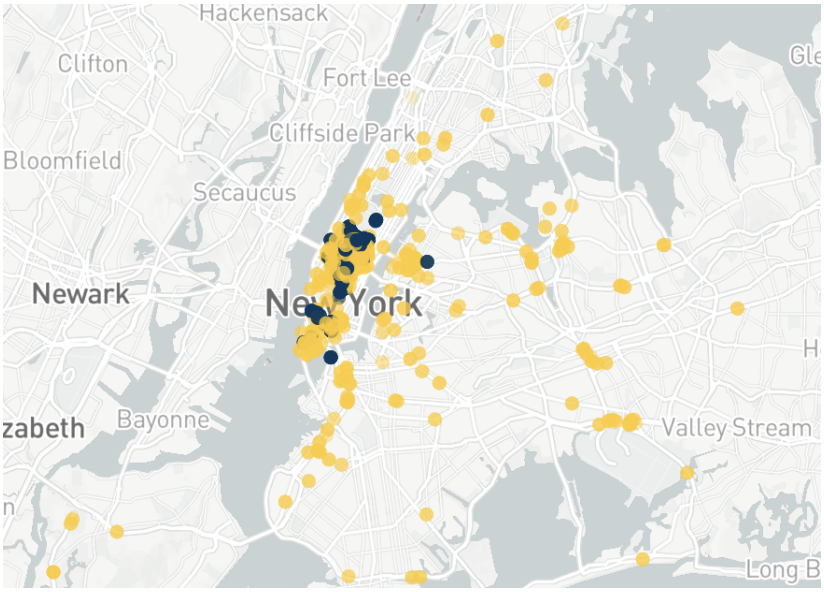


Figure 2 shows that the mean price steadily increases for each increment in number of stars and that the variance is highest for five-star hotels. The yellow dots in the figure denote outliers which, in this case, occur most frequently for three-star hotels. The same positive correlation remains if the star variable is substituted with hotel review score, see Appendix figure 8.

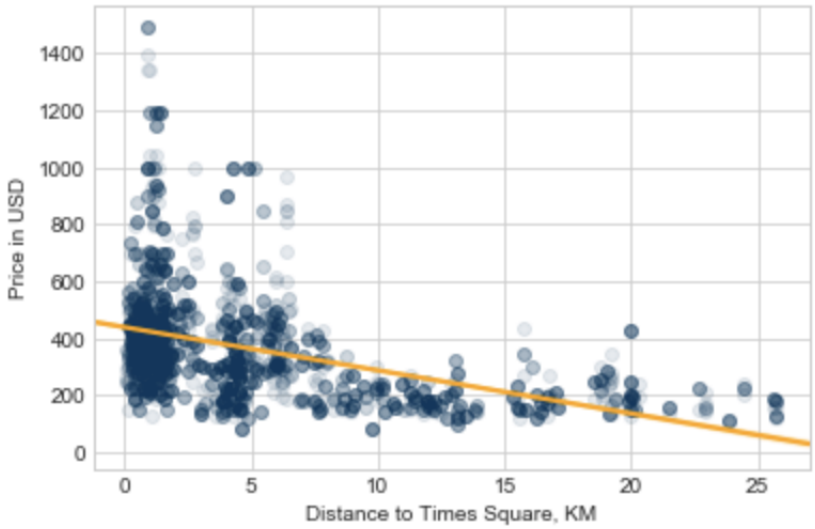
Figure 3: Mapped hotel prices: Most expensive percentile



Note: This map was made using the geographical plotting library plotly which plots each hotel room based on its coordinates. Each dot is an observation; the blue dots represent rooms belonging to the top 10 percent most expensive rooms. Only hotels in NYC's five boroughs are included on the map.

From figure 3, it is clear that while our sample contains observations scattered across all five boroughs, the majority of rooms are based on Manhattan. Furthermore, Manhattan holds the largest concentration of expensive rooms, expressed in the blue dots indicating observations belonging to the top 10 percent most expensive rooms. Based on this information, Manhattan is used as reference dummy in the OLS-, Lasso- and Ridge regression models.

Figure 4: Price in USD and distance to Times Square



Plotting distance to Times Square against hotel price in figure 4 shows a negative correlation which corresponds to the visualization in figure 3 . This finding confirms that it is indeed sensible to define Times Square as the city center.

Table 1: Descriptive statistics for dependent variables

	Mean	**Std**	**Min**	**Max**
Star-rate	3.48	0.83	1.5	5
Review score	4.07	0.48	1.8	4.9
No. of reviews	1859	3060	5	25238
Amenities				
Free WiFi	0.89	0.31	0	1
Room service	0.45	0.5	0	1
Spa and wellness	0.74	0.44	0	1
Pet policy	0.39	0.49	0	1
Laundry service	0.77	0.42	0	1
Premium TV	0.69	0.46	0	1
Parking	0.83	0.38	0	1
Free breakfast	0.32	0.47	0	1
Free cancellation	0.38	0.49	0	1
Weekend day	0.43	0.5	0	1
Dist. to Times Square (km)	4.87	5.66	0.06	25.7
Boroughs				
Bronx	0.02	0.13	0	1
Brooklyn	0.09	0.29	0	1
Manhattan	0.71	0.46	0	1
Queens	0.17	0.38	0	1
Staten Island	0.01	0.11	0	1

Notes: The sample consists of 2849 values and has been restricted to only include hotel rooms relevant for this analysis.

Finally, table 1 summarizes all explanatory variables. With a mean value of 4.07, it appears that review rating on average is more generous than hotel star rating which has a mean value of 3.48. The top three most common amenities are free WiFi, parking and laundry service and the three most uncommon amenities are free breakfast, free cancellation and pet policy allowance. Manhattan accounts for 71 percent of all rooms, followed by Queens which only accounts for 17 percent; together, Brooklyn and Staten Island hold 10 percent of the rooms in the data.

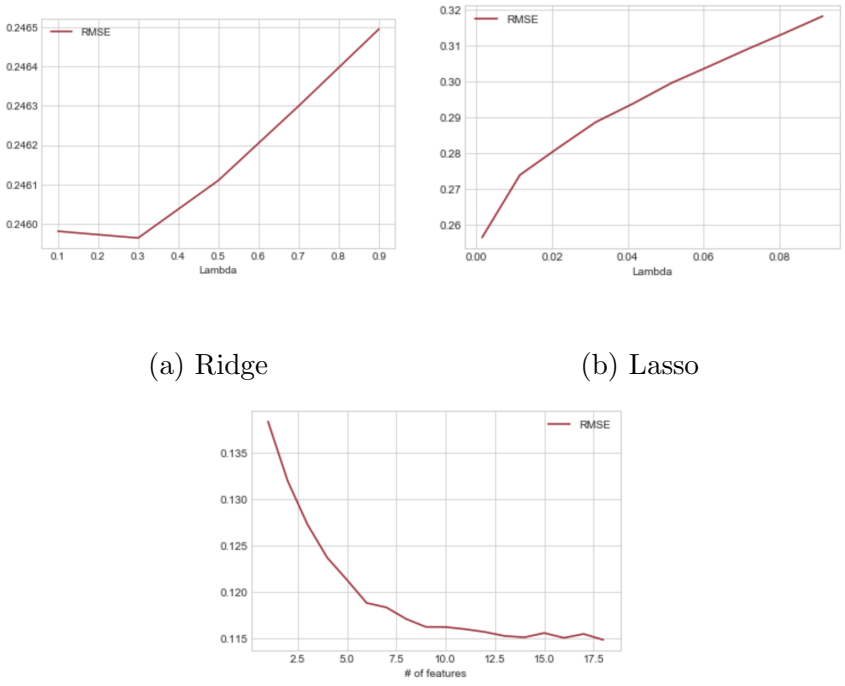
5 Empirical Analysis

Four different regression models have been used including the linear regression, Ridge regression, Lasso regression and the Random Forest regression. For all models, except the linear regression, the K-fold cross validation technique is implemented with 10 splits. Furthermore, the hyperparameter λ is applied in the Ridge and Lasso model. In the Random Forest model the RMSEs are measured given number of features put into the model. The Hold Out method is applied in all the regression models. The model with the lowest RMSE in the within-sample validation is the preferred model. The preferred model's generalized performance is tested on the test-subset of the data, and validated based on the learning- and validation curve.

5.1 Model Selection

The relationships between the hyperparameter value /number of features and the appurtenant RMSEs are depicted in figure 5.

Figure 5: RMSE in the regression models



(a) Ridge

(b) Lasso

(c) Random Forest

In table 2, the lowest RMSE, along with optimal value of the hyperparameter and number of features are depicted for each of the models.

Table 2: Optimal RMSE in the regression models

Model	OLS	Lasso	Ridge	RF
RMSE	0.2552	0.2563	0.2460	0.1148
λ	-	0.0015	0.3000	-
# of features	-	-	-	18

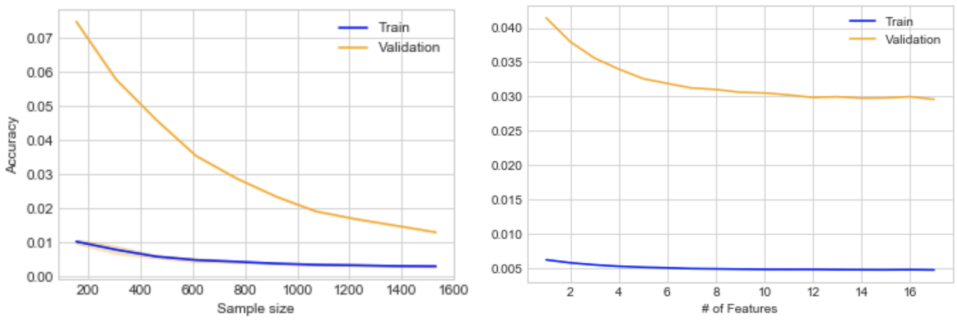
The linear regression model that has no hyperparameters as input returned a lowest RMSE of 0.2552. The RMSE for the Lasso and the Ridge model are fairly close, though the RMSE of 0.246 from the Lasso is slightly smaller than the RMSE of 0.256 from Ridge. This is likely due to Lasso's advantage of dropping features that turn out to be irrelevant, whereas Ridge only minimizes them. As seen in figure 5b, the optimal hyperparameter value of 0.0015 for the Lasso model is a corner solution. The corner solution reflects that a low degree of penalty is present in the optimal Lasso model. When considering at the coefficients for the optimal Lasso model, it can be identified that "Premium TV" variable has indeed been dropped (see Appendix table 5).
The Random Forest model has been performed without specifying the maximum for how deep each tree can be, as this is standard procedure. The lowest RMSE of 0.1148 obtained in Random Forest is returned by the optimal number of features of 18. It is clear that the Random Forest model is the preferred model as the RMSE value is significantly lower than those of the other models. Hence, the Random Forest model is the best in predicting the dependent variable.

5.2 Random Forest model validation

Since the Random Forest model has been selected as the preferred model, the next step is to validate the model. For this purpose, both a learning- and a validation curve have been estimated. But first, as the optimal number of features had been determined to be 18, the model's generalized performance is tested on test-subset of data, which yields a lowest RMSE value of 0.1153. Hence, the model preforms good on the test data as well.

Looking at the learning curve depicted in figure 6a, the gap between the validation- and training accuracy is quite large. As learning curve plots the accuracy of the model as the training set size increase, such a gap is an indication of high variance due to overfitting of the model.

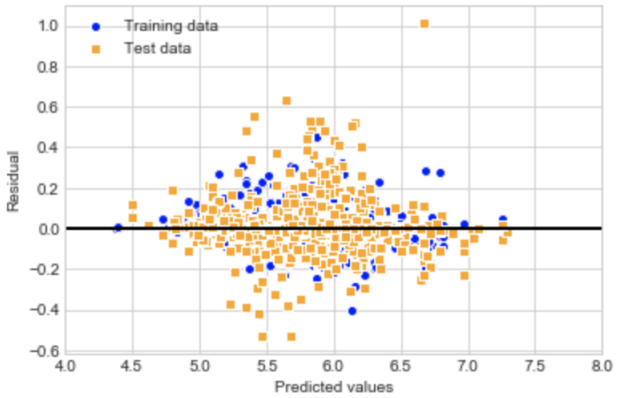
Figure 6: Model-fit investigation



(a) Learning curve (b) Validation curve

This problem of overfitting is also evident when considering the validation curve depicted in figure 6b. The two curves relate to one another, though the validation curve plots the number of features on the x-axis. As the models include more features, the accuracy of both the training and validation subset data increases, though the gap between the two curves does not seem to narrow.

Figure 7: Random Forest residuals



Next, the residuals from the Random Forest model are considered. These are shown in figure 7. In general, the model fits the training data better than the test data, which becomes evident through the yellow dot outliers. The residuals do not seem to be randomly distributed around the 0-line, which indicates that there is some information that the model is not able to capture.

Lastly, the importance of the explanatory variables is investigated. In table 3 the order of the features are shown with the most important features first. The feature importances are relative, why the sum of these equals to 1. The most important variables for predicting hotel room prices in this model are star-rate, the dummy indicating whether the hotel is located on Manhattan and the distance to Times Square.

Table 3: Random forest feature importance

Variable	**Relative importance**
Star-rate	0.362
Manhattan	0.209
Dist. to Times Square	0.117
Review score	0.105
No. of reviews	0.085
Laundry service	0.022
Free cancellation	0.020
Premium TV	0.011
Room service	0.010
Pet policy	0.009
Weekend day	0.009
Spa and Wellness	0.009
Parking	0.008
Free WiFi	0.008
Free Breakfast	0.007
Queens	0.007
Brooklyn	0.004
Bronx	0.001
Staten Island	0.001

6 Discussion

The purpose of this section is to discuss the results and some of the assumptions that were made during the analysis.

The data for this analysis was collected by scraping Expedia for information on hotel rooms in New York City. The obvious advantage of scraping is that it allows us to collect structured information about all of Expedia's hotel listings relatively fast. A disadvantage of scraping, however, is low reproducibility of results. Data is often added or removed from websites which makes scraping results inconsistent over time, resulting in a problem of validity. Since hotel room prices fluctuate a lot, this is expected to be especially true in the case with scraping Expedia. However, while scraping as a method for data collection might not be reproducible in terms of yielding consistent results, it *is* reproducible in terms of design, i.e. in its ability to allow reproduction of each step in the data collection process (Munzert et al. (2014)). The python script that automates the scraping is attached in the Jupyter notebook file, which allows other researchers to follow the same procedure.
The final dataset has a sample size of 2838 hotel rooms in New York City, which is arguably a relatively small dataset. To increase the sample size, Expedia could be scraped for hotel room prices in multiple cities. However, we chose not to do so since the general price level for each city could be different, which would result in higher variance and ultimately hurt the overall model performance.

From the empirical analysis, it was found that Random Forest regression was the best model to predict the hotel room prices of NYC. This model has a number of parameters that need to be specified, and the specification of each individual parameter has a high impact on how the model performs. For this reason, a brief explanation some of the important choices that were made in relation to the random forest model specification process is necessary.

First of all, the number of estimators was set to 1000. This means that the Random Forest regression is trained on a 1000 decision trees with the RMSE as a measure to split each node. In choosing the correct number of estimators, there is a trade-off between

decreasing the variance and slowing down computing performance. 1000 estimators are a lot; however, with a relatively small dataset, running the model on the data did not take too long. Similarly, increasing the number of estimators may not have had much of an impact.

Another specification that needs to be considered in relation to Random Forest Regression is deciding the maximum number of features that should be included in the model. This paper does not specify any maximum number since, in order to get the lowest RMSE value, we iterate through each feature to find the optimal number of features. This iterative process is visualized in figure 5c, where the optimal number of features included is 18, i.e. all of the explanatory variables.

Finally, in the regression model we allowed for the depth of the tree to be as deep as possible. The deeper the tree, the smaller the bias will be. Likewise, the model returned significantly higher RMSE if the maximum depth was set to the recommended square root of the number of features, which in this model is $\sim$4. However, not limiting the depth of the tree increases the risk of overfitting as the model will also fit noisy data. The learning- and validation curves depicted in figure 6a and 6b showed clear signs of overfitting. One way of overcoming this issue is to increase the data sample (though this may not help if the data cease to be noisy).

7 Conclusion

The purpose of this paper was to identify the optimal regression model to predict the price of a hotel room in New York City and further, within the preferred model, to determine which features were most important in predicting the price. The data for the analysis was scraped from Expedia.com for all available hotel rooms for two adults in week 19, 2019. The sample used in the analysis consisted of prices on 2838 hotel rooms along with information on their star-rating, location, Expedia guest reviews and amenities. From the empirical analysis it was found that the Random Forest regression model was best at predicting hotel room prices in NYC. However, by further inspecting the model, it seemed to be overfitting the data.

Lastly, the relative feature importances of the features in the Random Forest regression model were examined. From this it was concluded that star-rating was the most important feature followed by location, reviews, and amenities. These findings are in line with results found by M. Collins (2006).

8 Appendix

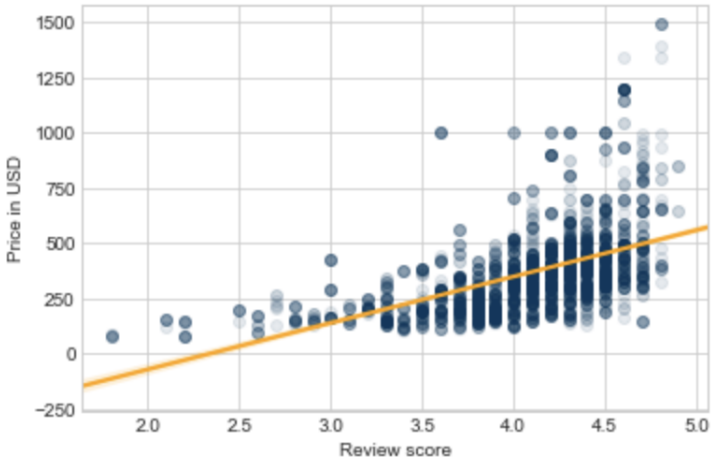
Table 4: Variable description

Variable	**Description**
Price in USD	Hotel price in USD
Star-rate	No. of stars for the specific hotel
Review score	Customers review on a scale from 1-5
No. of reviews	No. of reviews for the specific hotel
Free breakfast	Binary variable. Takes the value 1 if breakfast is free
Free cancellation	Binary variable. Takes value 1 if hotel has free cancellation policy
Premium TV	Binary variable. Takes value 1 if room has premium TV.
Laundry service	Binary variable. Takes value 1 if the hotel has laundry service.
Spa and wellness	Binary variable. Takes value 1 if the hotel has spa and wellness.
Room service	Binary variable. Takes value 1 if the hotel has room service.
Free WiFi	Binary variable. Takes value 1 if the hotel has free WiFi.
Parking	Binary variable. Takes the value 1 if parking is available
Pet policy	Binary variable. Takes the value 1 if pets is allowed to the hotel room
Weekend	Binary variable. Takes the value 1 friday-sunday
Distance to Times Square	Displays the distance (KM) to Times Square
Manhattan	Binary variable. Takes value 1 if hotel located on Manhattan.
Bronx	Binary variable. Takes value 1 if hotel located on Bronx.
Brooklyn	Binary variable. Takes value 1 if hotel located on Brooklyn.
Queens	Binary variable. Takes value 1 if hotel located on Queens.
Staten Island	Binary variable. Takes value 1 if hotel located on Staten Island.

Table 5: Lasso coefficients

Variable	Coefficient
Star-rate	0.2038
Review score	0.1580
No. of reviews	-1.3423
Free breakfast	0.0675
Free cancellation	0.1015
Free WiFi	-0.0032
Parking	0.01141
Pet policy	0.0732
Spa and Wellness	0.0262
Room service	0.0519
Laundry service	0.1192
Premium TV	0.0
Weekend	-0.0067
Distance to Times Square	-0.0146
Bronx	-0.0920
Brooklyn	-0.1981
Queens	-0.2357
Staten Island	-0.1837

Figure 8: Relationship between price and review score



References

BJH Advisors, BAE Urban Economics, VHB. 2017. NYC Hotel Market Analysis: Existing Conditions and 10-Year Outlook. Tech. rept. New York City Department of City Planning.

Lockyer, Tim. 2005. The perceived importance of price as one hotel selection dimension. Tourism Management, 529–537.

M. Collins, H.G. Parsa. 2006. Pricing strategies to maximize revenues in the lodging industry. International Journal of Hospitality Management, 91–107.

Munzert, Simon, Rubba, Christian, MeiÇner, Peter, & Nyhuis, Dominic. 2014. Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. 1st edn. Chap. PREFACE.

Raschka, Sebastian. 2015. Python Machine Learning. 1 edn. Chap. 10.

Rothstein, Ethan. 2018. The Worst Is Over For New York City's Rock-Solid Hotel Market. Bisnow.

Statista. 2017. Hotel services U.S. travelers are willing to pay extra charges for in 2017. https://www.statista.com/statistics/718206/hotel-services-us-travelers-are-willing-to-pay-extra-charges-for/. Accessed: 2018-23-08.

Wooldridge, Jeffery M. 2012. Introductory Econometrics. 5 edn. Chap. 3.