**Not Suite Dreams: An Analysis of Hotel Bookings and the Tendency for Patrons to Cancel**

Jacob Redmon, [jredmon@bellarmine.edu](mailto:jredmon@bellarmine.edu)

Christina Porter, [cporter@bellarmine.edu](mailto:cporter@bellarmine.edu)

**ABSTRACT**

Up to 150 word summary of your project.

1. **INTRODUCTION**

This project uses the k-nearest neighbor machine learning model to determine what factors, if any, can be used to anticipate the cancellation of a hotel booking by a patron.

1. **BACKGROUND**
   1. *Data Set Description*

The data set we used was data on hotel bookings from two hotels in Portugal, “originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.” It, “contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.”

We chose to use this data set because of its nice spread of categorical and numerical columns. The data was originally in two data sets, one for the resort hotel and the other for the city hotel. The original use of the data was for research in revenue management, machine learning, and data mining. The cleaned version of the data can be found here: <https://www.kaggle.com/jessemostipak/hotel-booking-demand>

* 1. *Machine Learning Model*

Provide one or two paragraphs that describe the ML models you considered for your data set and why you chose your final model. For your final model, explain in narrative how the model works. You do not have to provide a large amount of mathematical, but you can if want.

1. **EXPLORATORY ANALYSIS**

This data set contains 102894 samples with 31 columns, one of which we removed due to lack of data.

Some peculiar findings: the minimum value in the ‘adr’ column (Average Daily Rate) is -6, which must be an entry error as the rest of the row looks normal, the longest booking in advance was 629 days, and the date of arrival was more likely to be the 30th or 31st of any given month. Also, there is not much correlation in the data by looking at the heatmap of the correlation matrix in Figure 1. The most correlated columns were ‘stays\_in\_weekend\_nights’, ‘stays\_in\_week\_nights’, which is unsurprising. There was nothing notable in the histograms for the numerical data.

Shape

Description automatically generated

**Figure 1: Heatmap of the correlation matrix for the dataset’s categorical variables**

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| hotel | object |
| is\_canceled | int64 |
| lead\_time | int64 |
| arrival\_date\_year | int64 |
| arrival\_date\_month | object |
| arrival\_date\_week\_number | int64 |
| arrival\_date\_day\_of\_month | int64 |
| stays\_in\_weekend\_nights | int64 |
| stays\_in\_week\_nights | int64 |
| adults | int64 |
| children | float64 |
| babies | int64 |
| meal | object |
| country | object |
| market\_segment | object |
| distribution\_channel | object |
| is\_repeated\_guest | int64 |
| previous\_cancellations | int64 |
| previous\_bookings\_not\_canceled | int64 |
| reserved\_room\_type | object |
| assigned\_room\_type | object |
| booking\_changes | int64 |
| deposit\_type | object |
| agent | float64 |
| days\_in\_waiting\_list | int64 |
| customer\_type | object |
| adr | float64 |
| required\_car\_parking\_spaces | int64 |
| total\_of\_special\_requests | int64 |
| reservation\_status | object |
| reservation\_status\_date | object |

1. **METHODS**

In this section, describe how you prepared the data for your model and performed multiple experiments using different parameters for the model.

* 1. *Data Preparation*

Describe how you prepared the data for your model. For example, you might need to normalize the data, so variables with wider ranges of values don’t overshadow variables with smaller ranges. If you decide to drop variables from the model or create variables from existing columns, explain the process and the reasoning behind those decisions.

* 1. *Experimental Design*

You will run your model several times with different parameters to see what different results you get. In a table, describe your experimental parameters. Three or four experiments are sufficient. This is where you will describe how you divided your data into train, validate and test data sets. For example:

Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | All four (4) raw features with 80/10/10 split for train, validate, and test |
| 2 | All four (4) normalized features with 80/10/10 split for train, validate, and test |
| 3 | All four (4) raw features with 70/15/15 split for train, validate, and test |
| 4 | All four (4) normalized features with 70/15/15 split for train, validate, and test |
| 5 | All four (4) features with a square root transform on displacement, horsepower, weight and acceleration with 70/15/15 split for train, validate, and test. |
| 6 | All four (4) features with a square root transform on displacement, horsepower and weight with a70/15/15 split for train, validate, and test. |
| 7 | All four (4) continuous features and the three (3) categorical features with a square root transform on displacement, horsepower, and weight with an 80/10/10 split for train, validate, test. |

* 1. *Tools Used*

Describe all of the software tools you used to perform your data preparation and model implementation. For example:

The following tools were used for this analysis: Python v3.5.2 running the Anaconda 4.3.22 environment for Apple Macintosh computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 0.18.1, Numpy 1.11.3, Matplotlib 1.5.3, Seaborn 0.7.1, SKLearn 0.18.1, and Patsy 0.41. Provide a brief explanation of why you chose these tools.

1. **RESULTS**
   1. *Classification Measures*

Provide the classification measures for each experiment. For example, you could provide a contingency table for each model to measure how well it classifies data. You could also do an ROC curve (using SciKit Learn).

* 1. *Discussion of Results*

Discuss which of your models provided the best classification (or some other outcome if not classification). Explain why you think your best model was the best and why your worst model was the worst.

* 1. *Problems Encountered*

No project goes perfectly smooth. Discuss any problems you had with obtaining the data, preparing the data, implementing the model, or evaluating the model. **It would be highly unusual to indicate that you had not problems.**

* 1. *Limitations of Implementation*

Discuss the limitations of your model. Is there is reason it might not be the best way to model the data? What other models might work better?

* 1. *Improvements/Future Work*

What would you like to do to improve your model in future work? Do more experiments, use a different model, add/remove variables, find a different data set, etc?

1. **CONCLUSION**

Finish up with a paragraph or two of summarizing your problem, the results and your conclusions (good model, bad model, needs more work, etc.).

**REFERENCES**

Nuno Antonio, Ana de Almeida, Luis Nunes, *Hotel booking demand datasets*, Data in Brief, Volume 22, 2019, Pages 41-49, ISSN 2352-3409, <https://doi.org/10.1016/j.dib.2018.11.126>. *(https://www.sciencedirect.com/science/article/pii/S2352340918315191)*

**Other directions:**

1. 10-pt, Times New Roman, 1” margins all around (if you use this template you are already set).
2. Ensure all tables and figures are numbered appropriately and referenced in the text. See examples above and below.

|  |  |
| --- | --- |
| **Figure 1: Comparison of X/Y from dataset (single plot) (8 pt.)** | **Figure 2: (a) Function Output (b) A against B (multiple plots) (8 pt.)** |