Team: *Albert Z. Guo, Keith Pallo, Rhett Dsouza and Lukas Justen*

Status Report - Airbnb Price Evaluator

1.) Start by succinctly defining your task in terms of its inputs and outputs. For example, "our task is to predict the outcomes of baseball games based on attributes of the two teams, along with other factors such as weather and location." Then, briefly say why your task is important.

Given an excerpt of Airbnb listings in different cities of the US - our goal is to predict the best US-Dollar price per night for a given new apartment based on diverse attributes about the host and the apartment itself. The impact of our task is tremendous since we can utilize our model to help hosts find an acceptable price for their apartment. Also, this model can be helpful for guests who want to check if a given listing is overvalued or undervalued - particularly for those new to the platform, since private individuals are less likely to have any experience in the travel or hotel industry, which can make the cold start and price estimation particularly difficult for them. However, our tool can also be used by experienced Airbnb hosts who want to check if a specific renovation of their apartment (for example, adding a bath) justifies a price increase for prospective future guests.

2.) Describe the data set you have utilized to date. What types of attributes are there, how many attributes, how many examples, and how have you partitioned the data for the purpose of development/training/validation/testing.

Currently, we have used official Airbnb data sets provided via Kaggle for the cities of Seattle, Boston and New York in order to increase the diversity of listings in our training data, and to cover the landscape of price differences in the US. For each city we have roughly the same amount of 3,000-4,000 examples, making our final data set about 10,000 total listings. Each of the individual data sets offer the same 92 attributes describing the listing, it's host, and if available, it's reviews. After cleaning the data with Trifacta, our 10,000 examples have no missing values and contain both categorical and continuous values. Furthermore, we have divided the project into three different phases in order to be able to draw better conclusions about the relevance of our applied machine learning algorithms:

I. First, we consider attributes about the host (e.g. superhost, average response time, listing count, verifications, acceptance rates, host since date) and about the listing (e.g. latitude, longitude, accommodates, bathrooms, amenities, minimum nights, cancellation policy) as part of our first stage. We chose these attributes (22 in total) for the beginning since we think that we will be able to predict the price without requiring review data. Moreover, this will help us to assess if these attributes are alone accurate enough to predict the price of a listing - or if we really need reviews, pictures or additional descriptions.

II. Since our hypothesis is that reviews are important in order to assess if the listing price is reasonable, we want to use the different types of reviews at the second stage. Here we combine the features of the first stage with the information about the reviews (e.g. reviews per month, first/last review, communication rating, location rating). Finally, we can draw conclusions about the importance of reviews and how it's related to the use of different machine learning algorithms.

III. The last stage for this project includes text analysis (i.e. sentiment analysis of the reviews) or image analysis with deep neural networks. Here we want to focus on feature engineering in order to find new predictive features about the reasonability of a listing's price. Due to time constraints, this stage might be too much for our timeframe. However, we plan to continue with our work on this project. Thus, this step is crucial in order to increase accuracy of our model. Moreover, this helps to get a better accuracy on the cold start of a listing that has no reviews using a picture of the apartment.

In order to get a better insight of the final performance of our models we decided to set 10% of the data aside for a final evaluation of our models. In addition, we want to utilize 20% of the data during validation time and the remaining 70% of the data for the final training time. All in all, this will give us a better estimate regarding the performance and relevance of the model we create.

3.) Present your preliminary results on the task. Which learning techniques have you tried, and how have they performed? Note**:** Mention which existing machine learning software packages, if any, you are utilizing.

Since our task is a value prediction problem (i.e. we are trying to predict the price per night for an apartment), we focused on regression algorithms. The current algorithm we utilized is the Random Forest Regressor - which we implemented using the scikit-learn package in Python. We measured the performance of the model using scikit-learn's score function that calculates the coefficient of determination (R^2) of the model (https://bit.ly/2BgVBCO) (1.0 is the best possible value). We have a training set score of 0.909 and test set score of around 0.627. Our entire dataset size is reduced to around 11,250 after cleaning (e.g. removing NaN values etc.) and we binarized (one-hot encoded) a subset of our initial features to increase the total feature number to more than 60. Additional feature engineering (i.e. deriving/constructing more features based on the features we have already) may be required to boost our algorithm performance. So far, we only tested Random Forest Regressor, but will also try other regression algorithms, and use the boosting ensemble method (with a several different regression algorithms) as well.

4.) Briefly (1-2 paragraphs) describe your plans for the remainder of the quarter, and list any questions or concerns you have.

Our plans are briefly illustrated in the second question. We currently have the basic structure of the project and we are about to finish the first major phase. What we need to do in the remainder of the quarter is firstly try as many reasonable algorithms as possible to achieve a good prediction. From there we plan to explain and reason why certain algorithms perform better than the others, and detail how we have tuned the hyperparameters. Finally, we will set up the display website making sure to review and address reasonable critiques from peer reviews. If time allows, we also plan to utilize some of the popular deep learning techniques to analyze the comments of Airbnb users (e.g. numerical ratings generation based on user sentiment analysis; there are many open-sourced packages we can use) and the images posted by the hosts.