# Predicting "Deal Quality" for Airbnb Listings

Tu Hoang Cam Nguyen

Data 310

Report Part I

**Introduction**

The goal of this study was to create a predictive model that could find out the "Deal Quality" of Airbnb listings, which can be "Good," "Bad," or "Unknown." This report explains the steps that were taken to create a model that can make accurate predictions. These steps include cleaning and preparing the data, feature engineering, model selection, and evaluation.

## **Methodology**

**Data Cleaning and Preparation**

The initial step involved loading the dataset and performing necessary cleaning operations to ensure data quality. This included:

A computer screen shot of a program code

Description automatically generated

A screen shot of a computer code

Description automatically generated

* **Handling Missing Values**: Missing entries in columns like **name**, **host-name**, **last\_review**, and **reviews\_per\_month** were addressed by filling with placeholders or zeroes, as appropriate, to maintain dataset integrity without losing critical information.
* **Data Type Corrections**: The **last\_review** column was converted from a string to a datetime format to facilitate time-based calculations.
* **Dropping Irrelevant Columns**: Non-predictive columns such as **ID**, **name**, **host [id]**, and **host-name** were removed from the dataset, focusing the analysis on features likely to influence the outcome.

## **Feature Engineering**

A computer screen shot of a program code

Description automatically generated

To capture complex patterns within the data, several derived attributes were introduced:

* **Days Since Last Review**: Calculated to gauge listing activity and appeal over time.
* **Review Frequency**: Designed to measure engagement by comparing the total number of reviews against the duration listings have been active.
* **Price-Quality Ratio**: A metric intended to assess the balance between cost and perceived value, derived from the listing price and average review score.

These attributes aimed to enrich the dataset with nuanced insights, potentially correlating with "Deal Quality".

### **Model Development and Evaluation**

A screenshot of a computer program

Description automatically generated

**Model Selection**

A Random Forest Classifier was picked because it is reliable and can deal with complicated and non-linear connections between features. Its ensemble method, which combines the results of several decision trees, lowers the chance of overfitting and makes it easier to apply to new data.

**Training and Testing**

The dataset was divided into two sets: a training set with 80% of the data and a testing set with the rest of the data. The training set was used to teach the model, and for the first test, the hyperparameters were set to their default values.

**Evaluation Metrics**

The model's performance was assessed using accuracy and a detailed classification report, including precision, recall, and F1-score for each "Deal Quality" category.

## **Results and Interpretation**

The Random Forest model was right about 99.94% of the time on the test set, with almost perfect precision, recall, and F1-scores in all categories. This outstanding performance suggests that the mix of natural and engineered features accurately captures the underlying patterns that determine "Deal Quality."

It's likely that the derived attributes, especially those that show how recently something was reviewed (days since last review) and how well cost and quality were balanced (price quality ratio), were very important in getting such good predictions.

**Conclusion**

The method used in this study—which included cleaning up the data, engineering features, and carefully choosing which models to use—was very good at predicting the "Deal Quality" of Airbnb listings. The success shows how important it is to carefully choose features and a strong modelling approach when dealing with tough classification problems.