AIRBNB DYNAMIC PRICING RECOMMENDATION ENGINE

Links: Dataset | Tableau Dashboard

Introduction: Airbnb is an American rental company founded in 2008. The platform connects property owners with customers looking for a short or a long-term stay. It offers budget-friendly lodging, unlike the hotels, which can be very expensive for travellers. Pricing was a challenge for the company at the start, then they discovered dynamic pricing, a strategy that helps determine prices based on multiple factors.[1] Dynamic pricing involves many steps, like historical data and regression models. These models understand how different factors, like location and reviews, affect the price. The objective of this project is to use Airbnb historical data to determine the pricing predictors for the listings. This information will be further used to predict optimal prices for the properties by creating a linear regression model. The dataset chosen is Airbnb LA data. The dataset has 45886 rows and 18 columns. The attributes provide information about the multiple properties based in Californian cities, with the main emphasis being on Los Angeles. It has features such as host id, neighbourhood/city, room type, price, number of reviews, etc. The data goes all the way back to 2011. This project will be focusing only on the tuples from 2015 to 2025, which is the historical data of the last 10 years.

Data Dictionary.xlsx

Abstract: The project's purpose is to analyze price by city; room type and the number of reviews. The main objective is to create a regression model that helps us identify the price predictors. The first objective was met with the use of data visualization in Tableau. Bar graphs were created to find neighbourhoods and rooms with the highest average price. Correlation between two variables, price and number of reviews was checked using a scatter plot. I focused on the revenue generated by different cities, rooms and the reviews.

For the second, objective, a random regressor model was created to find feature importance. The top price predictors were obtained using the model with the help of feature importance. Furthermore, the model was used to predict prices based on the features of the property.

Tools Used: ***Jupyter notebook*** and ***Python*** were used for data cleaning, data pre-processing, to determine price predictors, feature importance and to create a random regressor model.
Tableau was used for analysis through data visualization. A dashboard was created showing price by reviews, room type and cities/neighbourhoods.

---

[1] (Shukla, 2024)

Steps Involved in Building the Project:

1. **Data Collection:** Searched on Kaggle for good data sources for data collection. Airbnb LA data was downloaded from the Airbnb website.
2. **Data Cleaning:** The rows for the years 2011 to 2014 were removed because my focus is on the last 10 years. Trivial columns like name and host name were deleted. The null values in the price column were replaced with median prices. The nulls in reviews per month were replaced with 0. Minimum_nights > 365 were capped to 365. Then, the duplicates were removed. This cleaned dataset was saved as a csv file for data visualization and analysis of how room type, reviews, etc. affect price.
3. **Feature Engineering:** Correlation was checked between variables using a heat map and box plot for categorical columns. Columns like, neighbourhood groups, few review columns were removed because they had low impact. I chose to keep latitude and longitude in the model for the location. To remove price outliers, I used quantile 99% to set an upper limit, values above the upper limit were dropped.
4. **Random Regressor Model**: Prices were transformed to log(1+x) of prices because prices varied widely across listings. This was done to prevent higher prices from overpowering the model. Feature importance was also generated. The correlation of determination of this model was generated, R squared is 0.62, this model based on log prices is a good fit. Then, a price engine was created at the end.

Conclusion: The first objective was to find the pricing predictors. The heatmap showing the correlation coefficients between variables shows that number of reviews and price have a correlation coefficient of -0.0036, **reviews do not affect price**. Feature importance of latitude (0.194923), longitude (0.18427), availability_365 (0.096577) and calculated_host_listings_count (0.084021), is the highest. **Hence, latitude, longitude, availability_365, and calculated_host_listings are are the best price predictors.** Latitude and Longitude have the highest feature importance. Hence, **location plays a huge role in determining price**.
Secondly, room_type_Entire home/apt has an importance of 0.303698. Other room types don't have much significance.

Furthermore, the data visualization in Tableau indicates that **LA listings have an average price of $261** and the other Californian cities excluding the incorporated ones have a measure of $424. **Bel- Air, a LA neighbourhood, has the highest average price**, it is second overall. First is El-Segundo city.
It also concludes that no. of reviews has no positive effect on price, they are negatively correlated with price. Therefore, **reviews are not important**.
Hotel room type has the highest average price of $5418.