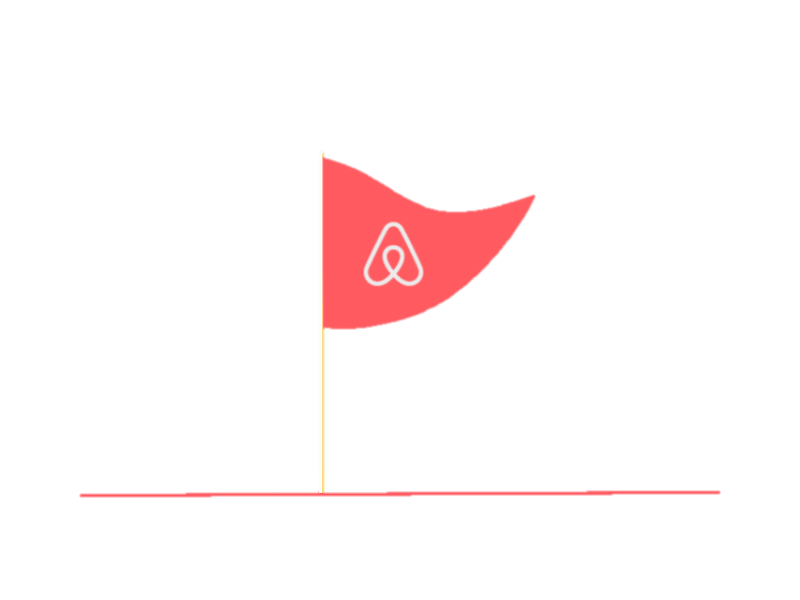
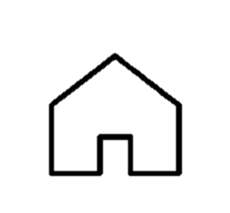
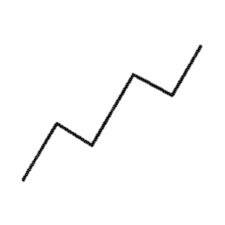
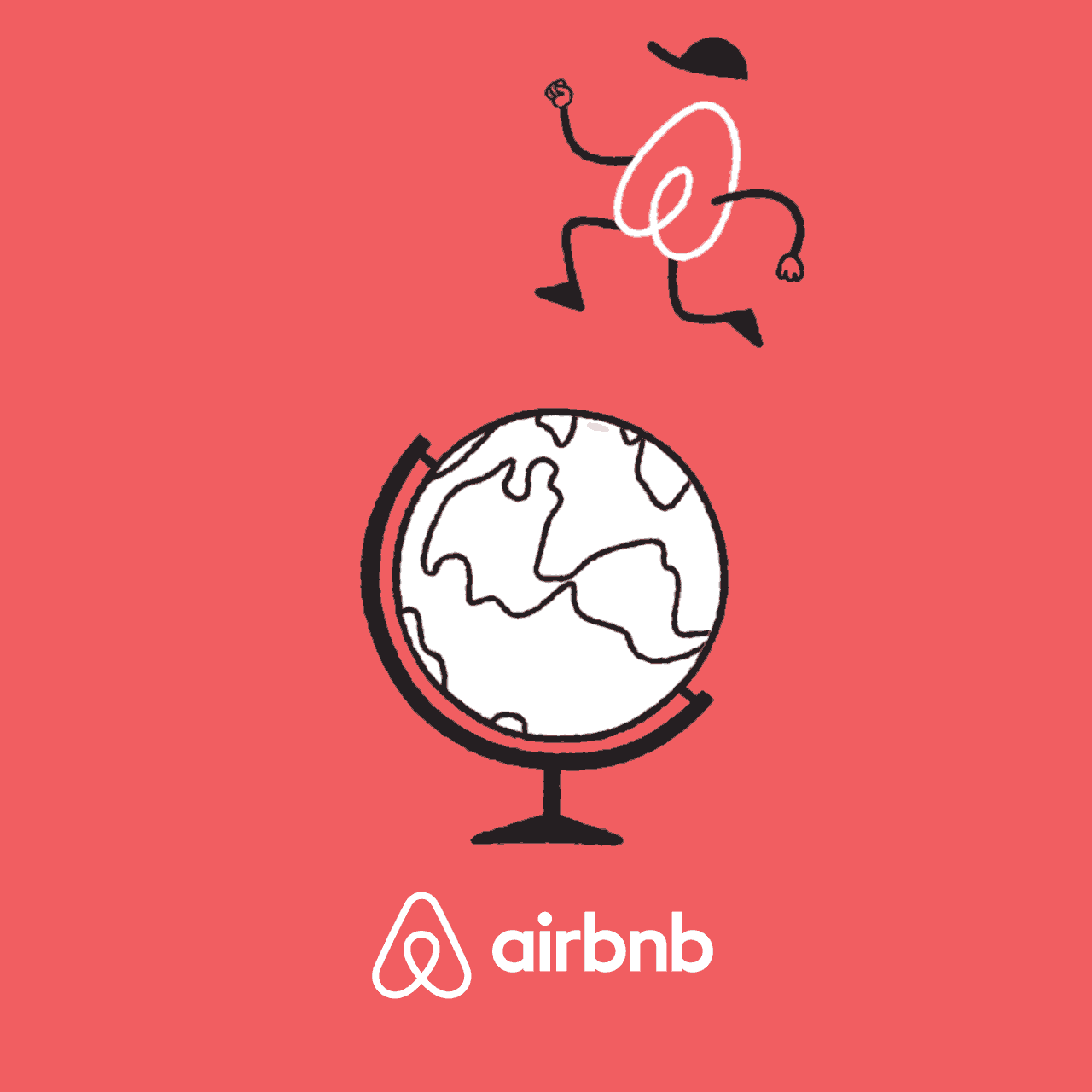
**Airbnb price prediction using Machine Learning**



**Avishay Aknin**

**Bar illan University**

**Data Science Course**

**Anatomy of an ML Project**

**1. Domain Knowledge-**

**Defining the Questions:**

* What factors most influence Airbnb property prices in Berlin?
* How accurately can we predict these prices given various features?

**Understanding the Problem:**

Airbnb property prices are influenced by factors such as location, property type, and amenities. Historical price data and property details are available for analysis, which will be used to understand these influences and build a predictive model.

**Outcome Definition:**

The goal is to create a predictive model with a maximum 10% error rate, enabling accurate price predictions for hosts and helping renters make informed choices. The model should also identify key factors impacting property prices**.**

**Identifying Influences:**

Key factors include response time, host status, location attributes (such as neighborhood and postal code), property features (including type, size, and room configuration), booking conditions, and review ratings.

**Exploring New Knowledge:**

Potential new information could include the experience level of hosts, such as their years of activity on the platform. Additionally, analyzing price distributions by categorizing properties into very high-end and budget-friendly segments could provide further insights into pricing dynamics.

**Project Design**

**Optimal Design Selection:**

To achieve our goals, we selected a combination of regression models, Random Forest, Support Vector Machines (SVM), and other advanced techniques. This mixed approach leverages the strengths of different models to enhance the accuracy of price predictions.

**Defining Research Subjects:**

We define and select research subjects by focusing on Airbnb listings in Berlin, analyzing various property features such as location, type, and amenities to identify what influences pricing.

**Data Handling Strategy:**

**Excluding Irrelevant Columns:**

* Removed columns that were deemed irrelevant for prediction due to their unique or unsuitable data types for the model.

**Filtering Location:**

* Eliminated all listings not located in Berlin to ensure the dataset is focused and relevant to the target area.

**Cleaning Data:**

* Addressed fields with anomalous values, such as properties with a price of zero, and corrected or removed these entries.

**Handling Missing Values:**

* Initially filled missing values using K-Nearest Neighbors (KNN) imputation.
* Removed over 10,000 records with incomplete or unusable data.
* Completed remaining missing values using Label Encoding to ensure data integrity and usability.

**Feature Selection:**

* Used various models to assess feature importance and selected those consistently deemed significant by multiple methods for the final dataset.

**Algorithm Validation:**

Validate algorithms through cross-validation and assess performance using metrics like RMSE or MAE for continuous price predictions to ensure accuracy and reliability.

**2.Data Preparation**

**Data Extraction**

Collect data from the Kaggle Airbnb Berlin dataset.

<https://www.kaggle.com/datasets/thedevastator/berlin-airbnb-ratings-how-hosts-measure-up>

**File Preparation**

Organize the data into structured flat files for consistency and ease of use.

### Data Preparation

1. **Clean Text Data:** Remove non-alphanumeric or numeric characters - (e.g., ;@#$%) from text fields to standardize and normalize the text.
2. **Reduce Large Categories:** Consolidate categorical variables with too many unique values into broader categories to simplify the dataset and enhance model performance.

### Exploratory Data Analysis (EDA) Steps:

1. **Basic Statistics:**
   * Calculate minimum, maximum, mean, and unique values for each column to get an initial understanding of the dataset.
2. **Data Overview:**
   * Use df.info() and df.describe() to understand data types, missing values, and summary statistics.
3. **Visual Analysis:**
   * Utilize the AUTOVIZ library to automatically generate visualizations and uncover patterns and trends in the data.
4. **Distribution Analysis:**
   * Examine the distributions of key columns, such as numeric columns such as price, to understand their ranges and variances.
5. **Sentiment Analysis:**
   * Analyze the sentiment of customer reviews to gauge overall customer satisfaction and its potential impact on the target variable.
6. **Price Distribution:**
   * Investigate the distribution of the price column, which is the target variable, to identify trends and potential outliers.
7. **Correlation Analysis:**
   * Evaluate the correlation between variables to understand relationships and dependencies, using correlation matrices or heatmaps.
8. **Geographic** **Analysis:**
   * Analyze geographic data by mapping property locations to identify patterns and trends based on location, using tools like Folium.

### Data Cleansing – Outliers and Missing Values

**Outlier Detection and Handling:**

1. **Check Distribution:**
   * Assess the distribution of data for skewness. Determine if data is left-skewed, right-skewed, or normally distributed.
2. **Normalize Using Z-Score:**
   * For data that is normally distributed, use Z-score normalization to identify and handle outliers. Values with Z-scores beyond ±2.5 are considered outliers.
3. **Apply IQR Method:**
   * For non-normally distributed data, use the Interquartile Range (IQR) method. Calculate IQR (Q3 - Q1) and define outliers as values outside 1.5 \* IQR below Q1 or above Q3.

### One-Hot Encoding and Label Encoding

**Label Encoding:**

* Applied label encoding to categorical columns where variables have a natural order or where the number of unique categories is high.

**One-Hot Encoding:**

* Used for the Room Type column, which has only two categories: "Private room" and "Shared room." This column was one-hot encoded to capture its impact on the price, representing each category with a separate binary column to avoid introducing ordinal relationships that do not exist.

### Feature Selection

1. **Apply Models:**
   * Used Lasso, SVM with L1 penalty, Gradient Boosting, and Random Forest to identify important features.
2. **Aggregate Results:**
   * Created a DataFrame to combine feature importance from each model.
3. **Select Features:**
   * Chose features selected by at least three models.
4. **Finalize Dataset:**
   * Constructed a final dataset with selected features and the target variable Price.

**Project Summary: Deployment and Beneficiaries of Machine Learning**

**How will we deploy the Machine Learning?**

The machine learning model will be deployed through an API that integrates with existing property management or booking platforms. This API will provide real-time price predictions and insights based on the model. The deployment process will include setting up cloud-based infrastructure for scalability, ensuring the model can handle high volumes of requests, and establishing data pipelines to continuously update the model with new data.

**Who will use and benefit from the Machine Learning?**

* Property Hosts: They will use the model to set competitive prices for their listings, optimizing their earnings while remaining attractive to potential guests.
* Renters: They will benefit from accurate price predictions, helping them make informed decisions and find better deals.
* Property Management Platforms: These platforms can integrate the model to offer value-added services, enhancing user experience and providing actionable insights for market trends and pricing strategies.

**Algorithms**

**Method Selection:**

For the Airbnb price prediction project, we chose supervised learning methods. Supervised learning is appropriate because we have labeled data (i.e., historical prices) and we aim to predict a continuous target variable (price). Algorithms like regression models, Random Forest, and Gradient Boosting are used to leverage this labeled data effectively.

**Model Validation Techniques:**

To validate model accuracy, we employ statistical tests such as Mean Squared Error (MSE) and R-squared. These metrics help in assessing how well the model predicts the price compared to the actual values.

**Dataset Partitioning**

We partition the dataset into training and testing sets, using a 70% training and 30% testing split. Additionally, we further divide the test set into two groups based on price categories: low-priced and high-priced apartments. This segmentation allows us to evaluate the model's performance separately for different price ranges, providing insights into how well the model predicts both ends of the price spectrum and ensuring that predictions are accurate across various price points.

**Boosting**:

We use Gradient Boosting to enhance model accuracy. It builds multiple decision trees sequentially, where each tree corrects the errors of the previous ones, improving prediction precision and handling complex patterns effectively.

**Hyperparameter Tuning:**

We used Grid Search to fine-tune hyperparameters. This method involves systematically testing various combinations of hyperparameters to identify the optimal settings that maximize model accuracy and minimize error.

**Implementation**

**Model Selection:**

We initially selected the best models based on their evaluation metrics. From these, we chose the top three performers for further refinement. We then optimized their hyperparameters and segmented the data into low and high price categories for more targeted predictions.

**Model Ensembling:**

We evaluated the top three models separately, examining each one's performance individually to determine the best one for accurate predictions.

**Testing on New Data:**  
In this project, we did not use completely new, unseen data for testing. Instead, we relied on cross-validation and split the existing data into training and testing sets to evaluate the models' performance.