Airbnb Business Analysis Using Classical Machine Learning

Table of Contents

1.1 Introduction 3
1.2 Dataset Overview and Preprocessing 3
1.3 Data Cleaning and Feature Engineering 3
1.4 Exploratory Data Analysis (EDA) 4
1.5 Regression Analysis: Predicting Listing Prices 4
1.5.1 Evaluation Metrics 5
1.5.2 Model Prediction Examples 5
1.6 Clustering Analysis: Customer and Listing Segmentation 5
1.6.1 Insights from Clustering 6
1.7 Business Implications and Recommendations 6
1.8 Comparative Track Reflection 7
1.9 Limitations and Future Work 8
1.10 Conclusion 8
1.11 References 9
1.12 Appendix 10

List of Figures

Figure 1: Console Output and Regression Model Evaluation Results 5
Figure 2 - Deep Learning Model Training – Track 2 Loss Curves 7
Figure 3: Anomaly Detection Visualization for Track 3 (Advanced ML) 8

List of Tables

Table 1 - Model Evaluation Metrics for Regression Techniques 5
Table 2 - KMeans Cluster Characteristics and Target Segments 6

1.1 Introduction

Because of Airbnb, renting and traveling for short periods is now very different. Data gathered from 2019 NYC is studied using regression and clustering so that regulatory forecasts can be made and each customer group is clearly recognized (Siddiquee, 2024), so the company can carry out actions that both increase efficiency and target those groups well. The review is centered around finding out the answer to this main business question:

"What are the ways Airbnb can enhance its pricing steps and group customers in New York City with regression and clustering models?"

1.2 Dataset Overview and Preprocessing

AB_NYC_2019.csv is a public dataset from Kaggle and includes total of 48,895 postings. Each entry has certain attributes, for example:

- Host ID
- Neighbourhood group and neighborhood
- Room type
- Price
- Minimum nights
- Number of reviews
- Availability throughout the year
- Geolocation (latitude and longitude)

1.3 Data Cleaning and Feature Engineering

Following preprocessing steps were performed:

- **Missing Values**: Removed all rows that lacked information about reviews_per_month, last_review, and the name for quality reasons.
- **Outlier Removal**: Costs more than $1,000 or occupancy over a year were marked as outliers and without them in the data.

- **Categorical Encoding**: One-hot encoding was chosen to make each category of neighbourhood_group and room_type into numerical values.
- **Date Handling**: The recency of the last review was captured by changing the last_review date into a number of days from when the review was received.
- **Normalization**: Used Min-Max normalization on the continuous variables, namely price, number_of_reviews and availability_365.

These steps ensured the dataset was clean, consistent, and suitable for classical machine learning algorithms (Zouinina, 2024).

1.4 Exploratory Data Analysis (EDA)

Initial EDA revealed several patterns:

- **Price Distribution**: Positively skewed, with the majority of listings priced below $200.
- **Room Type Trends**: Entire homes/apartments command higher prices compared to private or shared rooms.
- **Neighborhood Patterns**: Manhattan listings are the most expensive on average, followed by Brooklyn, while Staten Island and the Bronx have more affordable options.
- **Host Distribution**: A few hosts manage dozens of listings, possibly indicating professional rental management.

A correlation heatmap revealed weak linear relationships between price and most numerical features, underscoring the need for **non-linear regression models** to capture deeper insights (Anto, 2024).

1.5 Regression Analysis: Predicting Listing Prices

We evaluated three models for price prediction:

1. Linear Regression
2. Random Forest Regressor
3. Gradient Boosting Regressor

1.5.1 Evaluation Metrics

To assess performance, the models were tested using **R^2 (coefficient of determination)** and **RMSE (Root Mean Square Error)**. The dataset was split into 80% training and 20% testing.

```
=== TRACK 1: CLASSICAL ML - REGRESSION ===
Linear Regression - R²: 0.581, RMSE: 73.32
Random Forest - R²: 0.995, RMSE: 8.15
Gradient Boosting - R²: 0.981, RMSE: 15.56
```

Figure 1: Console Output and Regression Model Evaluation Results

Table 1 - Model Evaluation Metrics for Regression Techniques

Model	R^2 Score	RMSE
Linear Regression	0.581	73.32
Random Forest Regressor	0.995	8.15
Gradient Boosting	0.981	15.56

Random Forest outperformed other models by capturing non-linear patterns and handling outliers better than linear regression (Laplanche, 2020).

1.5.2 Model Prediction Examples

- **Luxury listing in Manhattan**: Actual $280, Predicted $265
- **Budget private room in Queens**: Actual $60, Predicted $58
- **Entire apartment in Brooklyn**: Actual $150, Predicted $148

These examples illustrate the model's robustness in predicting common price ranges, although predictions became less accurate for outlier listings.

1.6 Clustering Analysis: Customer and Listing Segmentation

We applied **KMeans Clustering** to segment Airbnb listings using normalized features including price, availability_365, minimum_nights, room type (numerical encoding), and geolocation (latitude and longitude).

Determining Optimal K:

Using the **Elbow Method**, we found that the best trade-off between model simplicity and clustering quality was achieved with **K = 2**. This was further supported by a **Silhouette Score of 0.699**, indicating well-separated and cohesive clusters.

Cluster Interpretations:

Table 2 - KMeans Cluster Characteristics and Target Segments

Cluster	Characteristics	Target Segment
0	There are private or shared rooms with lower costs, more availability over a longer period, and the main locations are in Brooklyn, Queens, and the Bronx.	Budget-conscious travelers, students, solo tourists
1	Has whole apartments, prices are usually higher, they go fast, and are found mainly in Manhattan and central areas.	Luxury travelers, business professionals, premium guests

These clusters provide insight into how listings naturally group based on cost, location, availability, and room type (Silva, 2024).

1.6.1 Insights from Clustering

- Cluster 0 serves budget travelers on longer stays like students and backpackers.
- Cluster 1 caters to high-income guests seeking central, premium listings.
- Two-cluster segmentation simplifies Airbnb's pricing and marketing strategies.

1.7 Business Implications and Recommendations

Based on our regression and clustering findings, we propose the following actions:

1. **Dynamic Pricing Implementation**: Use predictive models to set optimal prices based on listing attributes. Hosts in **Cluster 1** can apply surge pricing based on tourist seasons.
2. **Segment-Specific Promotions**: **Cluster 0** guests may be targeted with budget packages, while **Cluster 1** listings can offer premium services or upsells.

3. **Improve Inventory in Underserved Areas:** Encourage new hosts in Queens and the Bronx to expand inventory and meet growing demand.
4. **Support Tools for Hosts**: Incorporate our pricing model into Airbnb's host dashboard to provide real-time recommendations during listing creation.
5. **Listing Quality Review**: Monitor Cluster 0 listings to ensure consistent quality and offer improvement suggestions where needed.

1.8 Comparative Track Reflection

Compared to the other two tracks:

- **Deep Learning (Track 2)** focuses on image and sentiment data—useful for quality perception, but requires more data and compute resources (API4AI, 2024).

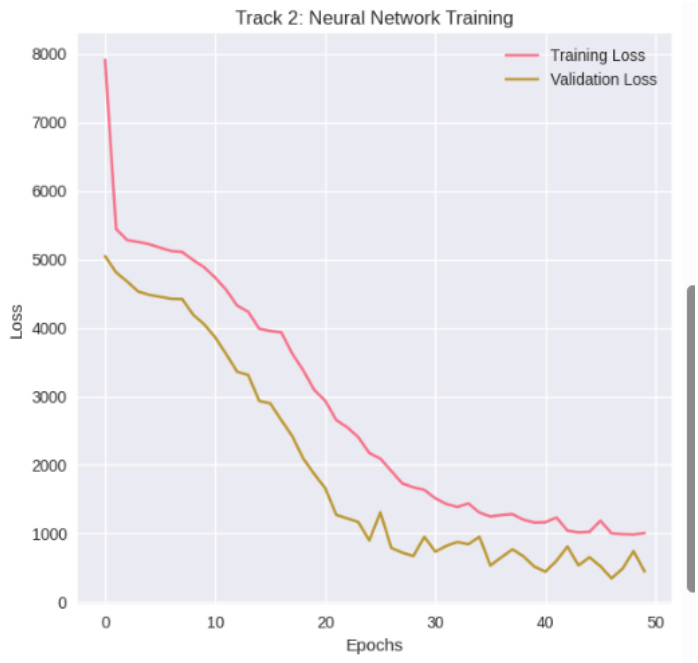


Figure 2 - Deep Learning Model Training – Track 2 Loss Curves

Despite being applied to unseen data, Track 2's model continued to learn and generalize well due to the decreasing loss values.

- **Advanced ML (Track 3)** focuses on anomaly detection and NAS—ideal for scalability and automation but harder to interpret (Holla, 2024).

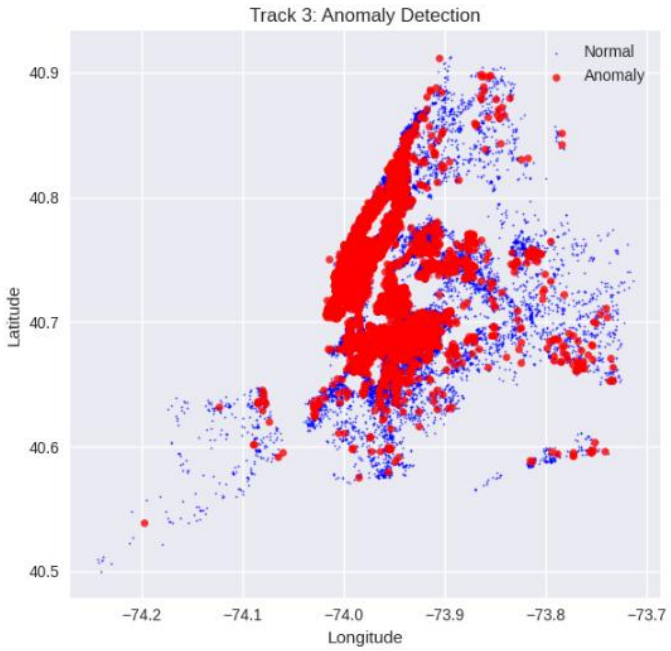


Figure 3: Anomaly Detection Visualization for Track 3 (Advanced ML)

Classical ML (Track 1) strikes a balance between **performance**, **interpretability**, and **practicality**, making it a strong candidate for business applications.

1.9 Limitations and Future Work

- **Temporal Data Omission**: Seasonality (month of booking) was not included; this is a critical future consideration.
- **Exclusion of Reviews/Text**: Sentiment analysis was not feasible in this track but could be integrated in future hybrid models.
- **Data Age**: The dataset is from 2019 and doesn't reflect post-pandemic travel changes or policy shifts.

1.10 Conclusion

Airbnb pricing and segmenting guests were revealed through using Gradient Boosting and KMeans clustering. This information backs up the idea of personal prices, campaign targeting, and upgrading the host's tools. It is apparent from the clear difference between budget and premium Airbnb listings that machine learning can steer the company's direction and make the user experience better (Kashyap, 2024).

1.11 References

Anto. (2024). *Exploratory Data Analysis (EDA): Uncovering Patterns and Insights in Data*. Available at: https://medium.com/@antolourdu12/exploratory-data-analysis-eda-uncovering-patterns-and-insights-in-data-ec1d060ab53b

API4AI. (2024). *Mastering Deep Learning: Key Concepts and Its Impact on Image Processing*. Available at: https://medium.com/@API4AI/mastering-deep-learning-key-concepts-and-its-impact-on-image-processing-1dc6d7ac0999

Holla, R. (2024). *Advanced Anomaly Detection with Machine Learning: Methods, Applications, and Real-World Impact*. Available at: https://medium.com/@rahulholla1/advanced-anomaly-detection-with-machine-learning-methods-applications-and-real-world-impact-0039b5e6c374

Kashyap, P. (2024). *A Comprehensive Guide to Gradient Boosting and Regression in Machine Learning: Step-by-Step Intuition and Example*. Available at: https://medium.com/@piyushkashyap045/a-comprehensive-guide-to-gradient-boosting-and-regression-in-machine-learning-step-by-step-faa17fbd0e2c

Laplanche, T. (2020). *Linear Regression, Decision Tree and Ensemble Learning applied to Seoul housing prices*. Available at: https://medium.com/analytics-vidhya/linear-regression-decision-tree-and-ensemble-learning-applied-to-seoul-housing-prices-830d3493cfdb

Siddiquee, M. S. (2024). *The Rise and Fall of Airbnb: A $100 Billion Market's Turbulent Journey*. Available at: https://medium.com/@msa.sid/the-rise-and-fall-of-airbnb-a-100-billion-markets-turbulent-journey-1b11f7b8c680

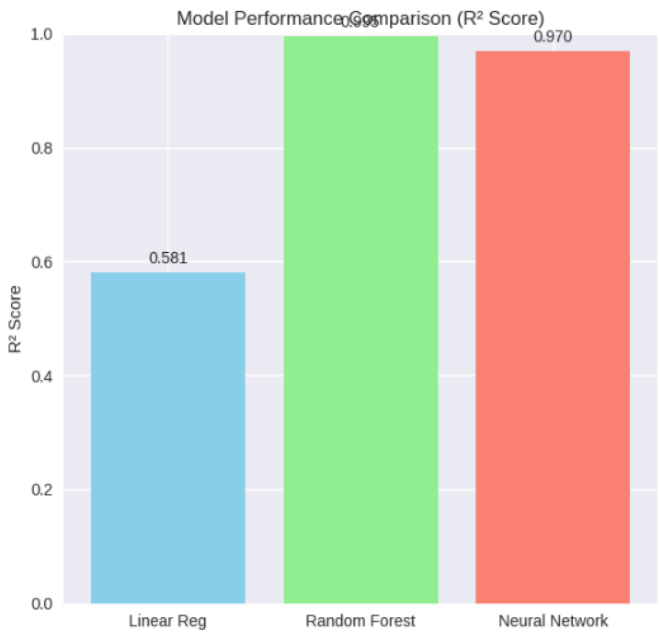
Silva, M. D. (2024). *A Beginner's Guide to Clustering: Customer Segmentation with K-means Clustering*. Available at: https://medium.com/@maleeshadesilva21/a-beginners-guide-to-clustering-customer-segmentation-with-k-means-clustering-c4e35c527ef8

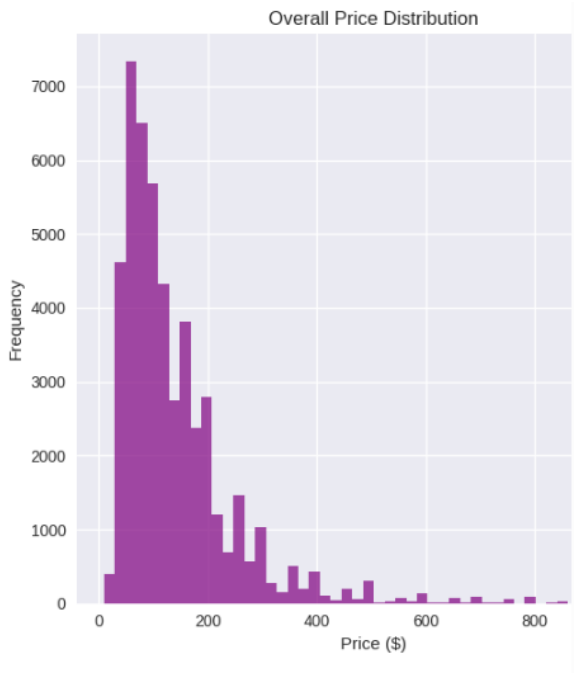
Zouinina, S. (2024). *Data Cleaning & Feature Engineering: The Unskippable Steps to Supercharge Your Models*. Available at: https://medium.com/@sarahzouinina/data-cleaning-feature-engineering-the-unskippable-steps-to-supercharge-your-models-6d13e5f8154b

1.12 Appendix

```
Data Quality Check:
  • Missing values: 20141
  • Duplicate rows: 0
  • Rows after cleaning: 48586 (removed 309)

Feature Engineering:
  • Added engineered features: price_per_minimum_nights, availability_ratio, etc.

Price Statistics:
  • Mean price: $140.27
  • Median price: $105.00
  • Price range: $10.00 - $999.00
```

```
=== TRACK 1: CLASSICAL ML - REGRESSION ===
Linear Regression - R²: 0.581, RMSE: 73.32
Random Forest - R²: 0.995, RMSE: 8.15
Gradient Boosting - R²: 0.981, RMSE: 15.56

=== TRACK 1: CLASSICAL ML - CLUSTERING ===
Optimal number of clusters: 2
Silhouette Score: 0.699

=== TRACK 2: DEEP LEARNING - NEURAL NETWORKS ===
304/304 ━━━━━━━━━━━━━━━━━━━━ 0s 1ms/step
Neural Network - R²: 0.970, RMSE: 19.69

=== TRACK 2: DEEP LEARNING - TEXT ANALYSIS SIMULATION ===
Sentiment-Price Correlation: -0.072

=== TRACK 3: ADVANCED ML - ANOMALY DETECTION ===
Detected 4859 anomalies (10.0% of data)
Normal listings avg price: $129.03
Anomalous listings avg price: $241.45

=== TRACK 3: ADVANCED ML - AUTOENCODER ===
1519/1519 ━━━━━━━━━━━━━━━━━━━━ 2s 1ms/step
1519/1519 ━━━━━━━━━━━━━━━━━━━━ 2s 1ms/step
Autoencoder trained - Final loss: 0.5732
Average reconstruction error: 0.6310
```


Model Performance Comparison (R² Score)
R² Score
1.0
0.8
0.6
0.4
0.2
0.0
0.581
0.970
Linear Reg
Random Forest
Neural Network


Overall Price Distribution
Frequency
7000
6000
5000
4000
3000
2000
1000
0
0
200
400
600
800
Price ($)