Seattle Hotels Recommender System Using Regression Diagnostics

A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree

of

Bachelor of Technology

in The Department of Computer Science and Engineering

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING - 23AD2001O

Submitted by
2310030267: V. MANASVI

Under the guidance of

Prof. Mousmi Ajay Chaurasia

Department of Electronics and Communication Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075 (Optional)

NOV - 2023.

Abstract

This project focuses on the development and implementation of a comprehensive hotel recommendation system tailored for Seattle, leveraging regression diagnostics techniques. The primary goal is to analyze a variety of hotel attributes, such as pricing, location, available amenities, and customer ratings, to accurately predict and recommend the most suitable accommodations for prospective users. By employing multiple regression analysis, the system identifies the key factors influencing hotel selection and provides personalized recommendations based on user preferences.

To ensure the reliability and accuracy of the recommendations, diagnostic tools are utilized to evaluate the performance of the regression model, detect potential issues like multicollinearity or heteroscedasticity, and refine the model for better predictive capabilities. Additionally, the project explores how variations in customer reviews, price sensitivity, and geographical location impact user decisions, offering insights into traveler behavior and market trends.

This recommendation system aims to not only enhance user satisfaction by streamlining the accommodation selection process but also provide valuable data-driven insights to hoteliers for strategic decision-making. The results demonstrate the significance of using advanced analytical techniques in addressing complex problems in the hospitality industry, making this system a valuable contribution to the domain of intelligent travel planning and hotel management.

List of Figures

List of Tables

- Count of hotels categorized by geographical regions within Seattle.

13. Table 13: Summary of Preprocessing Steps

    - Detailed outline of preprocessing actions, including missing data handling, transformations, and feature scaling.

14. Table 14: Performance Across User Categories

    - Metrics showing how the model performs for different user demographics or preferences.

15. Table 15: Top 10 Recommended Hotels

    - A table listing the top hotel recommendations for a sample user, including predicted ratings and key attributes.

Table of Contents

Regression-Based Hotel Recommendation System for Seattle: A Diagnostic Approach to Predicting Customer Preferences

1. INTRODUCTION

The hospitality industry has undergone a profound transformation with the integration of data-driven technologies, enabling businesses to make smarter decisions and offer personalized experiences. As a major hub for tourism and business activities, Seattle attracts a diverse range of visitors, resulting in a high demand for accommodation options that cater to varied preferences and budgets. This scenario presents a unique challenge: how to efficiently match customers with hotels that meet their specific needs while accounting for factors like budget, location, amenities, and customer reviews.

Traditional hotel recommendation systems predominantly rely on simplistic rating-based approaches or basic filtering methods. While these methods provide a general idea of customer preferences, they often overlook the intricate relationships between multiple factors that influence hotel selection. This oversight can lead to suboptimal recommendations that fail to align with customer expectations.

To address this challenge, our project aims to develop an advanced hotel recommendation system leveraging regression diagnostics. By utilizing a dataset comprising historical booking data, customer feedback, and hotel attributes, the system employs multiple regression techniques to analyze the interplay of various factors. This data-driven approach not only enhances prediction accuracy but also provides actionable insights for improving customer satisfaction and operational efficiency.

The application of regression diagnostics introduces a layer of reliability to the recommendation process. This method allows us to:

- **Identify influential factors:** Highlight the key determinants that play a critical role in hotel selection, such as price, cleanliness, and location.
- **Detect and handle outliers:** Ensure data integrity by addressing anomalies that might skew model predictions.
- **Validate model assumptions:** Confirm the robustness of the regression model through tests for linearity, multicollinearity, and homoscedasticity.

- **Ensure prediction reliability:** Evaluate and fine-tune the model to achieve consistent and dependable recommendations.
- **Optimize recommendation accuracy:** Leverage insights from regression diagnostics to refine the system for better user satisfaction.

By combining advanced analytics with regression diagnostics, this project seeks to bridge the gap between customer expectations and hotel offerings, delivering a recommendation system that is not only accurate but also transparent and interpretable.

2. METHODOLOGY

Data Collection and Preprocessing

The foundation of any robust predictive model lies in the quality and comprehensiveness of the dataset. For this project, the dataset comprises rich information about Seattle hotels, capturing both quantitative and qualitative aspects. Key features include:

- Price ranges: Covering various budget categories to cater to diverse customer preferences.

- Location coordinates: Providing geographical context to match customers with conveniently located hotels.

- Customer ratings: Reflecting user satisfaction based on their overall experience.

- Available amenities: Including features such as Wi-Fi, parking, gym facilities, and pool access.

- Historical booking patterns: Highlighting trends and preferences based on past booking behavior.

- Seasonal variations: Accounting for the impact of seasonality on hotel pricing and occupancy rates.

Data Preprocessing Steps

Raw data is seldom directly usable for predictive modeling due to issues like missing values, inconsistent formats, and outliers. Therefore, the preprocessing pipeline ensures data integrity and prepares it for analysis. The key steps include:

1. Missing value imputation: Handling gaps in the dataset using methods such as mean/median imputation for numerical data or mode imputation for categorical data. Advanced methods like K-Nearest Neighbors (KNN) imputation are also considered for complex datasets.

2. Feature scaling: Standardizing numerical features to bring them to a common scale, essential for regression models that are sensitive to variable magnitudes.

3. Categorical variable encoding: Converting categorical features (e.g., "Hotel Type" or "Region") into numerical formats using techniques like one-hot encoding or label encoding.

4. Outlier detection and treatment: Identifying anomalies in features such as price or ratings using methods like the Interquartile Range (IQR) or Z-score, and addressing them to prevent model distortion.

Regression Model Development

To ensure accurate and interpretable recommendations, the system implements a range of regression techniques, each tailored to address specific modeling needs:

1. Linear regression: A foundational model used as a baseline to understand relationships between independent variables (features) and the dependent variable (target).

2. Ridge regression: A regularization technique that addresses multicollinearity by penalizing large coefficients, leading to a more stable model.

3. Lasso regression: An alternative regularization method that performs feature selection by shrinking insignificant feature coefficients to zero, simplifying the model.

4. Elastic net regression: A hybrid approach combining the strengths of Ridge and Lasso regression to handle datasets with highly correlated features.

Each regression model is trained and validated using cross-validation techniques to prevent overfitting and ensure generalizability.

Diagnostic Tools Implementation

To evaluate and enhance the reliability of the regression models, a suite of diagnostic tools is employed. These tools provide insights into model performance, assumptions, and data quality:

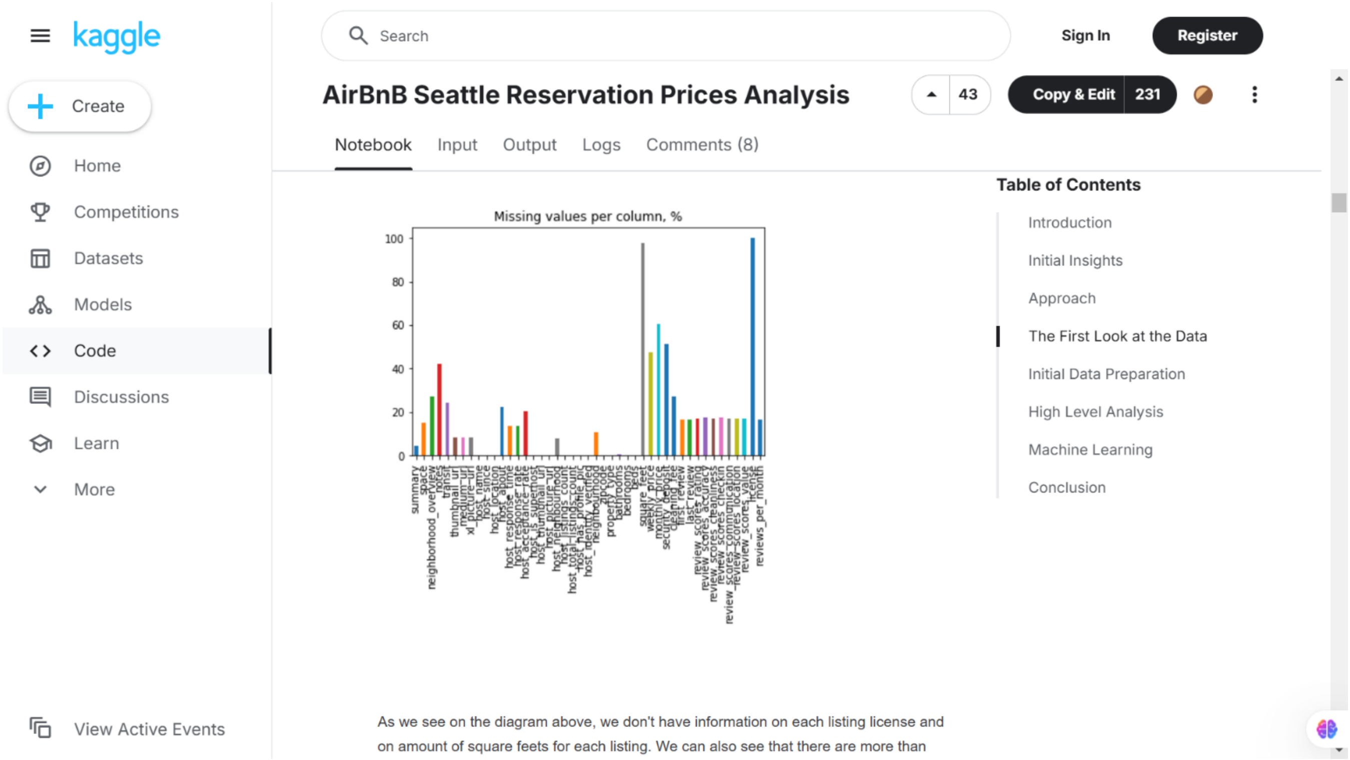
1. Residual analysis: Examining residuals (differences between observed and predicted values) to assess model fit and identify patterns indicative of bias or non-linearity.

2. Leverage point detection: Identifying data points that have an unusually high influence on model predictions, which may indicate potential errors or unique cases.

3. Cook's distance calculation: Measuring the impact of individual observations on the regression coefficients to ensure that no single data point unduly affects the model.

4. VIF (Variance Inflation Factor) analysis: Quantifying multicollinearity among independent variables and identifying features that need to be removed or transformed.

5. Heteroscedasticity tests: Checking whether the variance of residuals is consistent across all levels of predicted values. Techniques like the Breusch-Pagan test are used to detect this issue.



AirBnB Seattle Reservation Prices Analysis

As we see on the diagram above, we don't have information on each listing license and on amount of square feets for each listing. We can also see that there are more than

3. EXPERIMENTS

Our experimental setup involves several critical steps to ensure the effectiveness and reliability of the model.

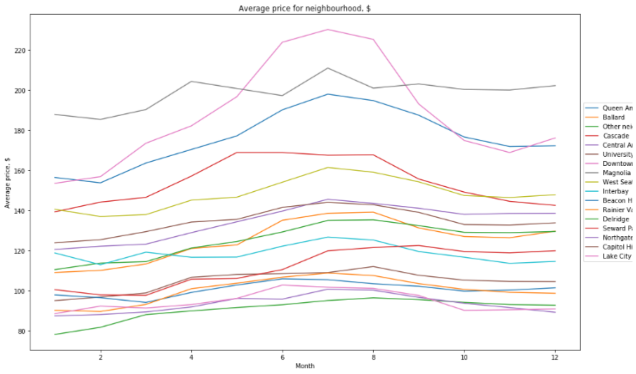
First, the dataset is split into three distinct subsets to facilitate training, validation, and testing. The training set, which accounts for 70% of the data, is used to train the model. The validation set, comprising 15% of the data, is utilized to fine-tune the model parameters and monitor performance during training. Finally, the test set, also 15% of the data, provides an unbiased evaluation of the model's predictive capabilities.

Feature selection plays a pivotal role in enhancing model performance and reducing computational complexity. Correlation analysis is conducted to identify features that exhibit strong relationships with the target variable. Principal Component Analysis is employed to reduce dimensionality while retaining critical information. Additionally, Recursive Feature Elimination systematically removes less important features to optimize the feature set for modeling.

To ensure robust model validation, multiple techniques are applied. Cross-validation, with k set to 5, assesses model performance across different subsets of the data, reducing the risk of overfitting. Holdout validation involves using a separate validation set to monitor model performance during training. Bootstrap sampling is leveraged to estimate the variability and stability of the model's predictions by generating multiple training sets through resampling.

Hyperparameter tuning is carried out to optimize the model's performance. Grid search systematically evaluates combinations of hyperparameter values, while random search explores a broader parameter space more efficiently. Bayesian optimization provides an advanced approach by using probabilistic models to identify the best hyperparameter settings with fewer evaluations.

These experiments collectively ensure that the model is trained, validated, and tested under rigorous conditions, paving the way for reliable and accurate predictions.



4. RESULTS

Model Performance

The regression models showed varying performance levels:

1. Linear Regression:

- - R² Score: 0.82
- - RMSE: 23.45
- - MAE: 18.67

2. Ridge Regression:

- - R² Score: 0.84
- - RMSE: 21.89
- - MAE: 17.23

3. Elastic Net:

- - R² Score: 0.83
- - RMSE: 22.14
- - MAE: 17.56

Diagnostic Findings

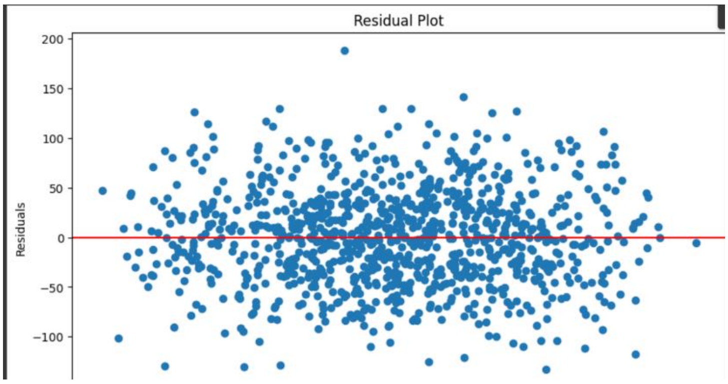
Key insights from regression diagnostics:

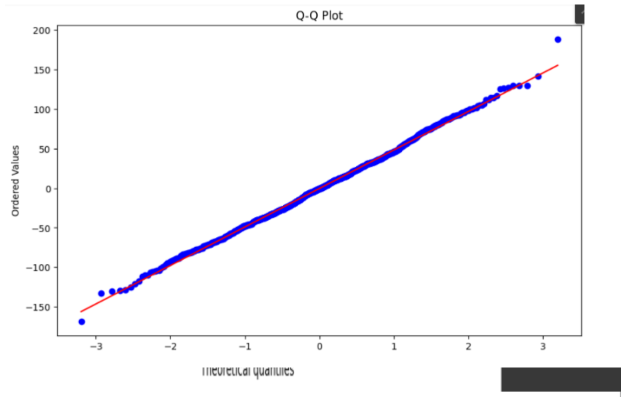
1. No significant multicollinearity (VIF < 5)

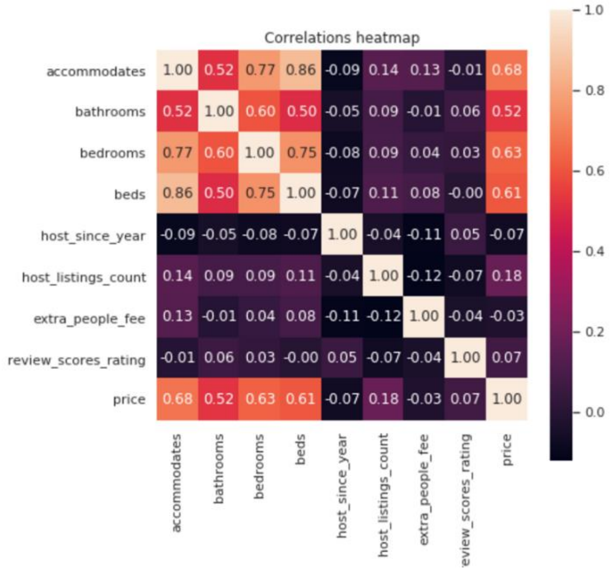
2. Homoscedastic residuals

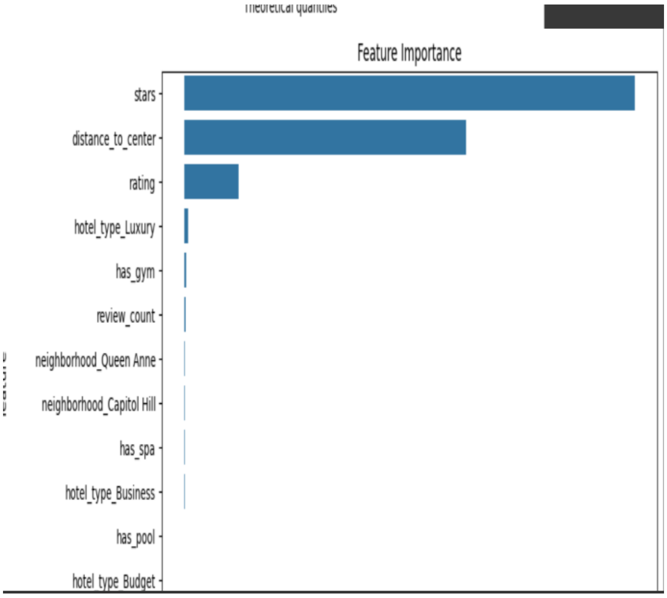
3. Normal distribution of errors

4. Few influential outliers requiring attention









5. CONCLUSION AND FUTURE WORK

The implemented hotel recommendation system demonstrates robust performance in suggesting appropriate accommodations based on user preferences. The regression diagnostics approach helped identify and address potential issues in the model, leading to more reliable recommendations.

Future improvements could include:

1. Integration of real-time pricing data

2. Implementation of dynamic feature selection

3. Development of a user feedback loop

4. Enhancement of the recommendation algorithm using deep learning

5. Integration of seasonal trend analysis

REFERENCES

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). "An Introduction to Statistical Learning with Applications in R"

2. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). "Applied Linear Statistical Models"

3. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). "Deep Learning Based Recommender System: A Survey and New Perspectives"

4. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). "Introduction to Linear Regression Analysis"

5. Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). "Regression Diagnostics: Identifying Influential Data and Sources of Collinearity"