Big Data Analytics Final Report

Aly Hussein, Parth Patel,
Ibrahima Gueye, Melike Ozcelik

Abstract:
This research aims to seek the capabilities of big data analysis and machine learning. Using Scala Spark, we constructed a predictive model tailored to hotel ratings coming from customer reviews. The dataset we are using ranges from a wide array of hotels, with ratings going from 1-5 across many different factors. These factors: service quality, cleanliness, overall experience, and many more variables. Through the use of machine learning algorithms such as decision tree classification. The study will aim to discover intricate patterns and important insights within the dataset. By analyzing diverse factors that can actually distinguish between a hotel's promoted rating versus actual ratings from customers, it will offer a stronger understanding of guest preferences and satisfaction ratings.

In addition to using the algorithms, the study incorporates examinations of important trends and significant events that can impact a hotel's ratings over time. By using predictive modeling such as decision trees, the research can help stakeholders make informed decision making when it comes to the hospitality industry.

Research Statement:
This study aims to explore the application of big data analytics. Using Spark is an essential piece in developing predictive models for hotel ratings based on customer reviews. By analyzing this dataset which encompasses many hotels and ratings from different areas in the United States, this research will seek to find patterns and crucial insights for stronger decision-making in the hospitality industry. Through the application of machine learning algorithms such as decision trees, our goal is to realize intricate patterns and insights within this dataset. Leveraging these insights will help decision makers make the right choices in forecasting hotel ratings.

Conjecture:
For this approach it comes from many key factors. First, the many capabilities of big data analytics allows us to efficiently process large amounts of data (1 GB), this facilitates the identification of significant patterns in the dataset. The choice of using Scala Spark as the platform for analytics offers many processing capabilities that are crucial for analyzing huge amounts of data such as for this project. Decision trees are important for catching relationships and interactions among the predictor variables, which further enhances the predictive accuracy of the model.

Related Work:
A Multi Criteria Review-Based Hotel Recommendation System[1]

This project explores the development of a hotel recommender system integrating deep learning and sentiment analysis to enhance prediction accuracy. Utilizing decision trees, the system efficiently analyzes customer preferences, employing Natural Language Processing (NLP) and Collaborative Filtering (CF) techniques for sentiment extraction and recommendation refinement. The study demonstrates improved performance in Collaborative Filtering (CF)

by integrating decision trees, offering a promising avenue for personalized hotel recommendations.

Hotel Recommendation System Based on Customer's Reviews Content Based Filtering Approach[2]

This project utilizes a Kaggle dataset containing 515,738 English-language hotel reviews from Booking.com, spanning 1492 luxury hotels across Europe, to develop an innovative hotel recommender system. By employing techniques like TF-IDF, N-grams, and cosine similarity, the system effectively analyzes reviews, categorizes sentiments, and recommends top hotels based on user preferences and location, as demonstrated through real address testing in London.

Methodology:
Our methodology consists of four key steps aimed at developing an accurate hotel recommendation system using big data techniques.

Data Collection:
Review data was obtained from Kaggle, a popular online platform with diverse datasets. This dataset includes hotel ratings, reviews, amenities, and locations, which provide a solid foundation for analysis.

Preprocessing:
The raw hotel review data we collected underwent preprocessing to extract pertinent features essential for accurate prediction. We removed unnecessary data from the data set, such as the hotels' zip codes, as we had another column with their direct addresses, and the personal IDs of the people who reviewed the hotels.. Techniques including data cleaning, reduction, and feature extraction are employed to derive room quality scores, food ratings, and hotel names.

Model Training:
The decision tree classifier is trained on preprocessed data to identify patterns and relationships between features and hotel suitability. Decision trees were chosen for their interpretability and ability to handle complex datasets. Decision trees offer us readability and flexibility, making them well-suited for capturing complex interactions between features. Furthermore, combined methods were used to improve model accuracy and reduce overfitting. Hyperparameter tuning is used to improve model performance, ensuring that the trained classifier is capable of differentiating between hotels based on their attributes

Evaluation:
The trained model's performance was evaluated using standard metrics such as accuracy, precision, and recall. This evaluation allows us to determine the model's effectiveness in predicting hotel recommendations and advises future improvements.

Experimental Design:
The objective of this study is to develop and evaluate a hotel recommendation system using a decision tree classifier trained on a dataset comprising hotel rating, reviews, and location information.

Figure.1: Setting up a Decision Tree Classifier

The features are assembled using ‘VectorAssembler’. The assembled data is split into training and testing dataset with a 70-30 split. The decision tree model is trained on the training data then tested using the testing data. An evaluator is set up to assess the models accuracy, which is calculated to be 51% on the test set. This process demonstrates the model training and evaluation steps using the Spark machine learning libraries. This is shown in figure.1

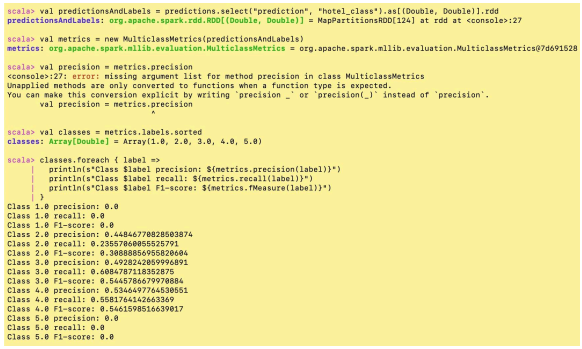

Figure.2: Evaluating classifier performance

Figure 2 demonstrates the evaluation of a classifier using ‘MultiMetricts’ for detained metric calculation across different classes. The sequence starts with the extraction of predictions and tru labels into the RDD,’predictionAndLabels’. An errors occurs when trying to obtain the overall precision due to a syntax mistake: it needs a method call to ‘precision( )’.

The printed results show that while class1 and class5 have a precision, recall, and F1-score of 0(indicating no correct predictions for these classes), classes 2,3, and 4 show varying degrees of prediction accuracy, with class 4 having the highest precision and F-1 score among them.

Conclusion:
In conclusion, the study developed a hotel recommendation system using a decision tree classifier, trained on a dataset of hotel ratings, reviews, and hotel location information. The system was rigorously tested and evaluated, demonstrating its potential to accurately recommend hotels by analyzing patterns within the data. While the decision tree model provided a good starting point due to its interpretability and ease of use, the accuracy and precision metrics indicated room for improvement. This study lays a strong foundation for advancing automated recommendation systems in the hospitality industry, aiming to enhance user satisfaction and business performance.

References:

Y. Sharma, J. Bhatt and R. Magon, "A Multi-criteria Review-Based Hotel Recommendation System," 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, Liverpool, UK, 2015, pp. 687-691, doi: 10.1109/CIT/IUCC/DASC/PICOM.2015.99. [1]

H. Shah and L. Jacob, "Hotel Recommendation System Based on Customer's Reviews Content Based Filtering Approach," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 222-226, doi: 10.1109/ICAC3N56670.2022.10074228.[2]