Enhancing Market Efficiency through Predictive Ratings: A Deep Dive into the Colombian Business Landscape

ABSTRACT

This paper explores the development and evaluation of predictive models for rating businesses in Columbia, focusing on the utilization of various machine learning techniques. Recognizing the profound impact of online reviews in consumer decision-making, as evidenced by a 2020 BrightLocal survey, this study aims to refine the predictive accuracy of business ratings using a dataset comprising 2,94,462 businesses with eight features. We emphasize the creation of new features from existing data, seeking to bridge the gap between actual and predicted ratings.

Initially, our study employs Logistic Regression and Random Forest models, selected for their respective strengths in binary classification and handling complex data sets. However, both models present limitations: Logistic Regression struggles with non-linear relationships and feature independence, while Random Forest grapples with interpretability and overfitting. To address these challenges and accommodate the diverse nature of our dataset, including textual data, we transition to a custom-built multi-input deep learning model using TensorFlow's Keras API. This approach is tailored to handle the dataset's complexity, particularly excelling in processing the high-dimensional and diverse data inherent in customer reviews.

Our model evaluation, incorporating a combination of 100 TF-IDF features and 22 newly created relevant features, reveals significant insights. Logistic Regression achieved an accuracy of 0.69, while Random Forest slightly outperformed with an accuracy of 0.71. The deep learning model, however, is expected to surpass these traditional methods by effectively capturing complex, non-linear relationships in the data.

The study contributes to the field by illustrating the effectiveness of deep learning in predictive rating systems, particularly in dealing with varied and high-dimensional datasets. It also highlights the importance of feature engineering and the selection of appropriate machine learning models based on the specific characteristics of the dataset. This research offers a comprehensive framework for businesses to leverage customer reviews for improved service quality and market positioning.

INTRODUCTION

Predictive ratings play a pivotal role in both enhancing market efficiency and ensuring the integrity of online review systems, essential in today's digital marketplace. These ratings serve as a critical tool for guiding consumers efficiently to high-quality businesses, crucial in an era where decision-making can be overwhelming due to the abundance of choices. According to a BrightLocal survey in 2020, a staggering 87% of consumers rely on online reviews for local businesses, underscoring the significance of these ratings in shaping consumer behavior. By condensing extensive user feedback into accessible metrics, predictive ratings not only streamline the decision-making process but also heighten consumer satisfaction by steering them away from potentially disappointing experiences.

Furthermore, predictive ratings stimulate a healthy competitive environment. Research in the Journal of Consumer Research (2019) indicates that businesses often improve their offerings in response to customer feedback and ratings. This competition drives an upward spiral in product and service quality, fostering innovation and superior customer service, which are integral to market growth and consumer satisfaction.

We have details about 2,94,462 of Columbia's businesses with 8 features. Most of the ratings given by users are 5 stars but we want to predict ratings by creating new features from the existing features and check the difference in predicted and actual ratings given to different businesses.

LITERATURE REVIEWS

In the quest to develop an effective predictive rating model for businesses in Columbia, we draw inspiration and knowledge from seminal works and research studies in the field of machine learning and data analytics. These works provide a foundational understanding of various methodologies and their applications, which are instrumental in guiding our approach.

Random Forest in Ecological Data Analysis:

Study by Cutler et al. (2007): This comprehensive research titled "Random forests for classification in ecology" provides an in-depth comparison of Random Forest with other machine learning algorithms. Published in the journal 'Ecology,' it highlights the effectiveness of Random Forest in handling complex datasets typically found in ecological studies. The study demonstrates the algorithm's superiority in terms of accuracy and interpretability, crucial factors in our selection of Random Forest for initial model building. The insights from this research are particularly relevant for our project, given the complexity and diversity of our dataset, which requires a robust algorithm capable of managing

high-dimensional data and providing interpretable results.

Application to Predictive Ratings:

Drawing from this study, we leverage the strengths of Random Forest in managing diverse features within our dataset, including both numerical and categorical data. The ability of Random Forest to provide a comprehensive view of feature importance and handle non-linear relationships is particularly valuable in our endeavor to predict business ratings accurately.

Foundations of Logistic Regression:

Work by Hosmer Jr et al. (2013): The book "Applied Logistic Regression" is a comprehensive resource that delves into the nuances of logistic regression. It discusses the theory, application, and interpretation of this statistical method across various fields, including medical, social, and behavioral sciences. The book's detailed exploration of logistic regression as a tool for binary classification is crucial for our project, especially in the early stages of model development.

Relevance to Rating Prediction:

In our project, we apply the principles outlined by Hosmer Jr and colleagues to model the binary aspects of customer ratings. Their work guides our understanding of how logistic regression can be effectively utilized in our context, despite its limitations in handling complex, non-linear data structures.

Hybrid Recommender Systems:

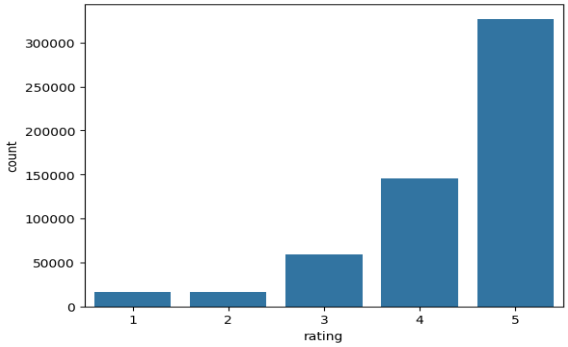
Survey by Burke (2007): Burke's extensive survey on hybrid recommender systems merges the strengths of collaborative and content-based approaches. His research sheds light on various hybridization techniques and their effectiveness in enhancing recommendation accuracy and diversity. This study forms a foundation for modern recommender systems, which aim to integrate multiple data sources and recommendation methodologies.

Implications for Our Project:

The insights from Burke’s research are particularly relevant for our task of feature engineering and model selection. By understanding the principles of hybrid recommender systems, we aim to create a model that not only predicts ratings accurately but also encapsulates the multifaceted nature of customer feedback and business attributes. The concept of hybridization informs our approach to combining different types of data — numerical, categorical, and textual — and developing a comprehensive model that captures the essence of consumer reviews and business features.

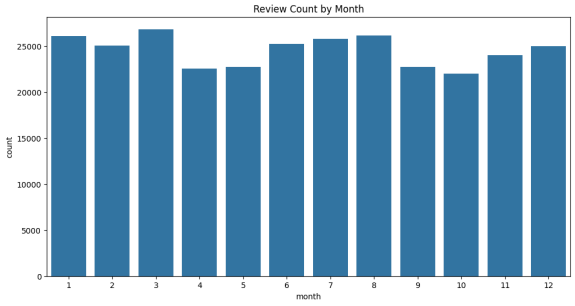
DESCRIPTIVE AND EXPLORATORY ANALYSIS

Bar Chart Visualization: We employed a bar chart to visualize the distribution of business review ratings, ranging from 1 to 5 stars. The y-axis of the chart represents the frequency of each rating. Notably, the data reveals a predominant occurrence of 5-star ratings, indicating a trend of positive feedback from consumers. This observation suggests high customer satisfaction but also hints at a potential skewness in the data towards positive ratings. Understanding this distribution is vital for our predictive modeling, as it influences how we handle and interpret the ratings data.

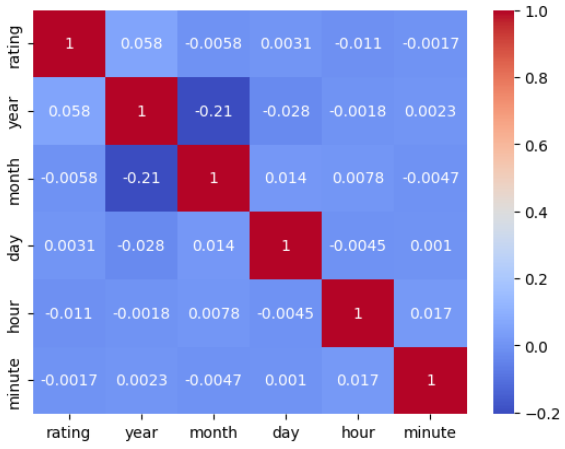


Content Analysis: To delve into the qualitative aspect of the reviews, we generated a word cloud from customer feedback. This word cloud highlights the most frequently mentioned terms in the reviews, with larger words representing higher frequency. Key terms such as "service," "staff," and "food" were prominent, along with positive adjectives like "great" and "delicious." This analysis provides valuable insights into the aspects that customers most frequently discuss and appreciate, such as service quality and culinary experience. Such insights are instrumental in understanding customer priorities and preferences.

Bar Chart for Monthly Trends: Another bar chart was used to depict the distribution of monthly review counts over a year. We observed significant peaks in January and December, with a noticeable dip around mid-year. This pattern suggests the presence of seasonal trends in customer review behavior, likely influenced by factors such as holidays or seasonal business activities. Recognizing these trends is crucial for our analysis, as it allows us to consider the temporal aspects of consumer behavior in our predictive modeling.



Heatmap Visualization: A heatmap was created to analyze the correlation coefficients between review ratings and various time-related variables, including year, month, day, hour, and minute. Interestingly, the ratings showed minimal correlation with the time of review. This finding indicates that while time-related factors are present, they do not significantly influence the review scores. This insight is essential in guiding our feature selection, emphasizing the need to explore variables beyond mere time-related factors to predict ratings effectively.



OBJECTIVE

In our current project, our primary goal is to develop a predictive model that accurately forecasts ratings based on a variety of features available in our dataset. To enhance the predictive power of our model, we are not just relying on existing data; we're also innovatively creating new features that we believe are pertinent to this task.

Our approach involves a meticulous analysis of the dataset to identify patterns and correlations that could be pivotal in predicting ratings. We plan to experiment with different combinations of features, including those newly created, to understand their impact on the model's predictive accuracy.

FEATURE ENGINEERING

Handling Text Data:
TF IDF Features: We have transformed the text data from reviews using TF IDF (Term Frequency-Inverse Document Frequency), which helps in capturing the importance of words in the reviews. TFIDF is beneficial for understanding the impact of specific terms on the ratings.

Review and Response Analysis:

Review Length: The length of each review was calculated, likely as a feature, as it can sometimes correlate with the sentiment of the review (longer reviews might be more detailed, either positively or negatively).
Response Length: Similarly, the length of the response (if present) was also added as a feature, indicating engagement or customer service quality.

Time of Day Classification:

We segmented the time of each review into distinct categories: 'Night', 'Morning', 'Afternoon', and 'Evening'. This classification aims to uncover possible patterns in customer behavior or satisfaction levels at different times.

For instance, reviews posted at night might differ in tone or content from those posted in the morning, reflecting various customer experiences or expectations.

Seasonal Analysis from Month Data:

We extracted seasonal information from the month data, categorizing each review into a corresponding season. This feature is crucial for identifying seasonal trends in customer feedback. For example, certain businesses might receive higher ratings during specific seasons due to seasonal product offerings or holiday-related services.

Weekday vs. Weekend:

We introduced a binary variable to differentiate reviews posted on weekdays from those on weekends. This distinction is significant as it could reveal variations in customer expectations and experiences between regular workdays and leisurely weekends, potentially influencing the nature and content of reviews.

Inclusion of Photos:

A binary variable was added to indicate whether a review includes a photo. The presence of a photo is often a strong marker of customer engagement and can be associated with more detailed or empathic feedback. Photos may also enhance the credibility of the review and provide visual evidence of customer experiences.

Feature Cleaning:

We streamlined the dataset by removing features such as 'gmap_id' and 'user_id', which were deemed irrelevant for our predictive modeling. This step ensures a cleaner, more focused dataset, enhancing the efficiency and accuracy of our model.

MODEL SELECTION

In our project, we initially chose Logistic Regression and Random Forest as our baseline models. We selected Logistic Regression for its effectiveness in binary classification problems and its interpretability. It's a robust model that performs well with categorical data, making it ideal for initial evaluations, especially when working with datasets that have binary or categorical output variables.

Logistic Regression has its limitations, including an assumption of linearity between variables, a need for feature independence, difficulty with complex output structures, and sensitivity to outliers.

Scalability issues: For this we have used a standard scaler to scale data from the columns of length of review, length of response because without scaling the model cannot interpret the data correctly.

Random Forest was another choice for a baseline model due to its ability to handle a large number of features and its robustness against overfitting. As an ensemble method that builds multiple decision trees and merges them for more accurate and stable predictions, it was ideal for getting a broad understanding of the feature importance and the initial predictive power of our dataset.
Random Forest, while robust, can be complex and less interpretable, may struggle with too many features, requires more resources for training, and can overfit noisy data.

Ultimately, we decided to use a deep learning model for our final analysis. Our decision was influenced by the model's ability to handle complex, non-linear relationships in data, especially given the varied nature of our dataset,

which includes numerical, categorical, and textual data. The deep learning model, particularly with its capacity to process textual information through layers like embeddings, offered a more nuanced understanding of the customer reviews. This was crucial for our objective of accurately predicting business ratings and deriving meaningful insights from the reviews.

Custom-Built Multi-Input Deep Learning Models using TensorFlow's Keras API, although powerful for complex tasks, tend to overfit on smaller datasets, demand large amounts of data, lack interpretability, require significant computational resources, and have long training times.
By comparing the results from the baseline models with those from the deep learning model, we aimed to explore different analytical angles and ensure the robustness of our final predictions.

MODEL EVALUATION

After adding 100 TF-IDF features and 22 relevant features we created using the dataset ['rating', 'text', 'pics', 'resp', 'review_length', 'len_of_response','is_weekend','season_1','season_2','season_3','season_4','weekday_0','weekday_1','weekday_2','weekday_3','weekday_4','weekday_5','weekday_6','time_of_day_Night','time_of_day_Morning','’time_of_day_Evening’,’time_of_day_Afternoon’ & TF-IDF features].

We have splitted data into train and test (70%, 30% respectively), fitted models based on training set and predicted values for test set to understand the accuracy of the model.

Comparison with Alternatives: Compared to simpler models such as logistic regression or random forests, this deep learning model likely excels in handling the high-dimensional and diverse data inherent in customer reviews. The effectiveness of this model, as opposed to the more traditional methods, would largely depend on its ability to capture complex, non-linear relationships in the data, which might be oversimplified by less sophisticated models.

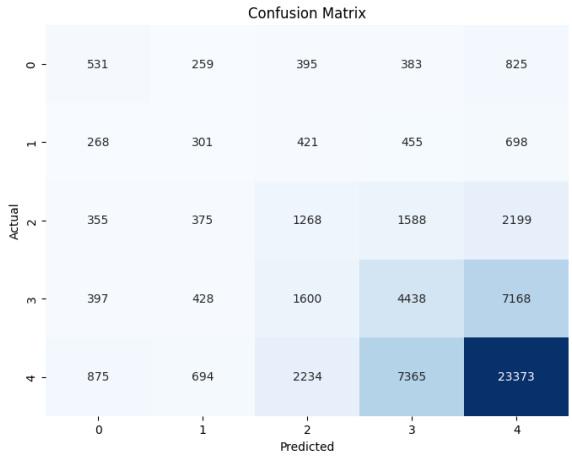
Feature Representations
Effective Features: The textual data, processed through an embedding layer, likely provided significant insights due to its ability to capture semantic meanings in customer reviews. Numerical features like review length and response length, and categorical features like time of day, likely contributed valuable dimensions to the model's understanding.

Model’s Parameters Interpretation

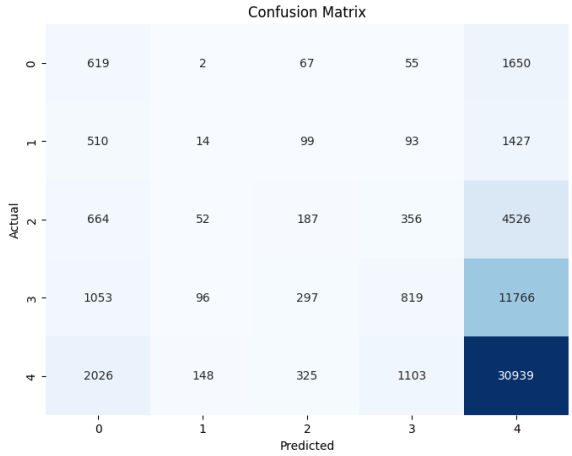
Deep Learning Parameters: The parameters in a deep learning model, especially in a complex architecture like this, are not as interpretable as those in simpler models like logistic regression. The weights in dense layers and embeddings represent abstract features learned from the data, which are difficult to translate directly into intuitive insights.

RESULTS AND CONCLUSION

Logistic regression has provided an accuracy of 0.69. The confusion matrix is as follows. The actual ratings are mentioned in the rows and predicted ratings by our model are mentioned in the columns.



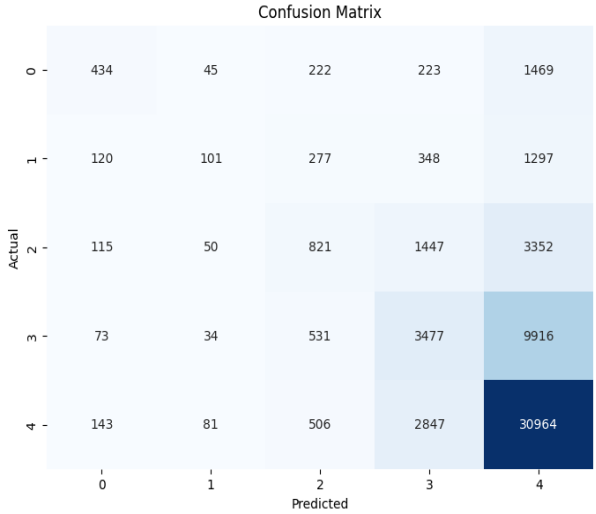
Similarly, Random forest classifier has provided the following results with the accuracy 0.71.



Multi-Input Model: Our model is designed to process multiple types of data inputs - numerical, categorical, and text data. This multi-input approach is tailored to handle the complexities and varied nature of the dataset effectively.

The model is compiled with the Adam optimizer and mean squared error as the loss function, which aligns with regression objectives.
The mean_absolute_error (mae) is used as a metric for model evaluation during training.
The training process involves 50 epochs with a batch size of 32, and a portion of the training data is used for validation (validation_split=0.3).

Model Evaluation: Post-training, the model's performance is evaluated on a test dataset to assess its predictive accuracy.



Literature Review References

- Random Forest in Ecological Data Analysis:
  Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792.

- Foundations of Logistic Regression:
  Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. John Wiley & Sons.

- Hybrid Recommender Systems:
  Burke, R. (2007). Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction, 12(4), 331-370.