Ed-Tech Apps Ratings Analysis

Problem Statement

The education technology (EdTech) sector has witnessed significant growth, revolutionizing the delivery of knowledge. However, understanding the factors influencing app ratings is crucial for further development. This project explores the shifts in educational app ratings and aims to uncover key insights.

Data Sources

- Datasets: Google Play Store Apps (kaggle.com)
- Backup Repository: https://github.com/gauthamp10/Google-Playstore-Dataset

Data Wrangling

- Removed duplicated records (478).
- Dropped columns with limited information on app ratings.
- Handled missing values in "Installs," "Size," and "Released" features.
- Excluded non-US Dollar pricing for non-free apps.
- Converted the "Free" feature to "Price," with 0 indicating free apps.

Pre-Processing

- Grouped continuous features into ordinal categories.
- Regrouped "Content Rating" into fewer nominal categories.
- Calculated new features such as "Age" and "Updated_age."
- Scaled and categorized raw rating scores.
- Handled boolean features like "Editors Choice," "In-App Purchase," and "Ad Supported."

Exploratory Data Analysis

- Explored feature distributions and target value distributions before and after transformation.
- Identified class imbalance in target values.
- Examined feature correlations, highlighting relationships between installations, in-app purchases, and app age.
- Uncovered insights into the impact of being on the Editor's Choice list on rating scores.

Modeling

Attempted multi-class classification but faced severe class imbalance.

- Conducted binary classification by combining ratings into "Low" and "High" classes.
- Employed classifiers (DecisionTreeClassifier, RandomForestClassifier, GradientBoostingClassifier).

- Achieved improved results with a macro-f1 score of 0.63 and an accuracy score of 0.74.
- Identified installation as the most crucial feature, followed by in-app purchase, app age, and updated age.

Project Structure

```
|-- data/
|   |-- raw_data.csv
|   |-- cleaned_data.csv
|-- notebooks/
|   |-- data_wrangling.ipynb
|   |-- exploratory_data_analysis.ipynb
|   |-- modeling.ipynb
|-- README.md
```

Current Questions

- **App Feature Significance:**

What features contribute significantly to determining app ratings, considering the unique nature of Ed-Tech apps? How have these features been refined through data cleaning and preprocessing?

- **Temporal Changes in App Landscape:**

How has the landscape of educational apps evolved over time, specifically focusing on changes in the number of apps, their ratings, and user engagement? What insights can be drawn regarding the overall growth and adaptability of Ed-Tech apps?

- **Effect of Imputed Data:**

In instances where data was imputed, such as in the case of missing "Released" dates, how has this affected the overall analysis and model outcomes? Are there discernible patterns in apps with imputed data compared to those with complete information?

- **Correlation Analysis:**

How does the correlation heatmap reveal relationships between app ratings and various features? Are there notable patterns, especially in terms of installations, in-app purchases, and app age, that significantly influence the rating scores?

- **Impact of Class Imbalance:**

How did the initial attempts at multi-class classification highlight the challenges posed by severe class imbalance? What led to the decision to combine rating scores into binary classes, and how has this adjustment affected the classification results and model performance?

- **Model Interpretability:**

For the selected GradientBoostingClassifier model, how do the feature importance scores shed light on the factors driving app ratings? Are there specific features that stand out, and how do they contribute to the overall classification of apps into "Low" and "High" rating categories?

Who Cares and Why

Understanding the factors influencing educational app ratings is essential for:

- **App Developers:**

Gain insights into features and characteristics that contribute to higher app ratings, improving the development process.

- **Educational Institutions:**

Identify trends that can enhance the selection of educational apps for students, optimizing learning experiences.

- **Investors and Stakeholders:**

Make informed decisions based on the performance trends of educational apps, aligning investments with market demands.

This analysis provides valuable information for key stakeholders, contributing to the improvement and strategic decision-making within the EdTech industry.