INFO 7390

ADVANCES IN DATA SCIENCES AND ARCHITECTURE

YELP-YODLERS

Ashritha Goramane	002728794
Rishab Dinesh Singh	002743830
Kshama Aditi Lethakula	002784433

Problem Statement

The challenge lies in leveraging Yelp data to gain insights into factors that contribute to the success of restaurants and predicting the success of new or existing restaurants based on various features and attributes

Goals

- Identify key factors that contribute to the success of business based on historical data from Yelp
- Predict the success of a business in a given location with a quantifiable measure of success, defined by the business's star rating and review count

Project Summary

- **Data Preprocessing**
  - Prepared the dataset by handling missing values, standardizing data, and detecting outliers.
- **Exploratory Data Analysis (EDA)**
  - Identified patterns, detected anomalies, and conducted hypothesis testing through visual and statistical methods.
- **Feature Engineering**
  - Enhanced features to improve predictive accuracy.
- **Model Development and Training**
  - Established a baseline with Linear Regression.
  - Implemented advanced models like Decision Trees, Random Forest, Gradient Boosting, and XGBoost.
  - Performance Evaluation: Used relevant metrics to evaluate and select the top-performing model.

Technical Stack

- **Programming language**
  - Python 3.11
- **Python packages**
  - pandas==1.5.3
  - scikit-learn
  - matplotlib
  - seaborn
  - plotly
  - nbformat
  - Xgboost

Conclusion

After comparing Linear Regression, Random Forest Regressor, and XGBoost, XGBoost (with learning_rate 0.5, max_depth=6) emerges as the preferred model for predicting business performance based on the provided dataset.

Reasons for Choosing XGBoost as final model

- Outperformed in predictive accuracy
- Robust compared to other models
- Optimized the model performance on hyperparameters

References

- https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html
- https://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_regression.html
- https://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_regression.html
- https://xgboost.readthedocs.io/en/stable/parameter.html
- https://www.geeksforgeeks.org/xgboost-for-regression/