Project Report

Title: Predict Restaurant Ratings

Subtitle: Leveraging Machine Learning for Enhanced Customer Satisfaction

Author : Haripriya R

Date: 24.04.2024

Internship at: Cognifyz Technologies

Abstract:

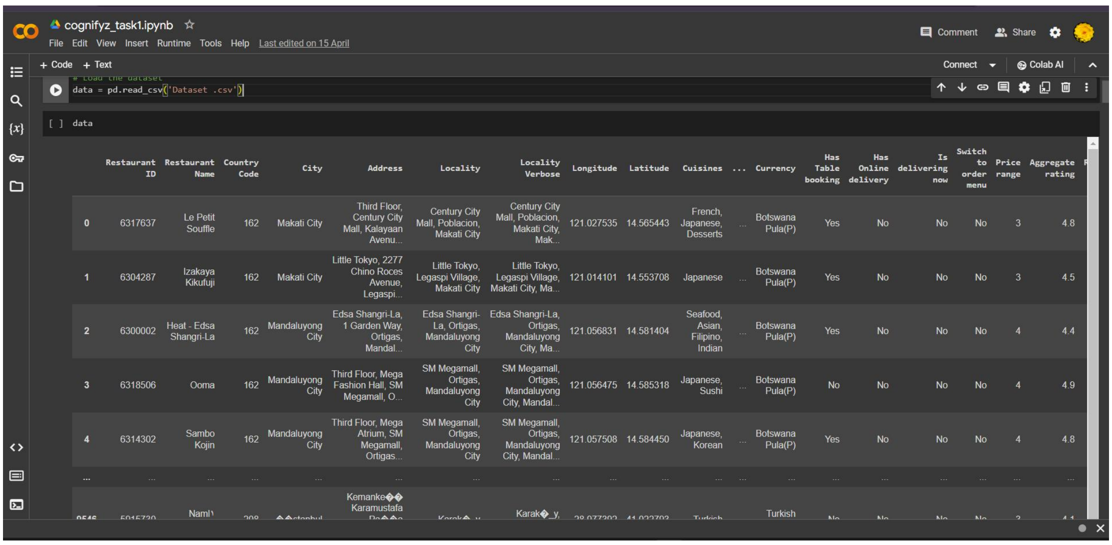
The project aimed to develop a machine learning model to predict the aggregate rating of restaurants based on various features. By leveraging data science techniques, the model seeks to provide valuable insights to restaurant owners and stakeholders for enhancing customer satisfaction and business performance. This report outlines the methodology, implementation details, results, and analysis of the project.

1. Introduction:

The introduction provides context for the project by discussing the growing significance of online reviews and ratings in shaping consumer decisions in the restaurant industry. It underscores the challenges faced by restaurant owners in maintaining high ratings amidst fierce competition and the increasing reliance on data driven approaches for strategic decision making. The introduction sets the stage for the project's objectives and highlights its potential impact on the restaurant industry ecosystem.

2. Data Preprocessing:

This section delves into the intricacies of data preprocessing, including data cleaning, feature engineering, and normalization. It discusses techniques for handling missing values, outliers, and categorical variables, ensuring the dataset is suitable for model training. The section also explores methods for addressing class imbalance and selecting relevant features to improve model performance. Visualizations such as correlation matrices and distribution plots are used to illustrate preprocessing steps and insights gained from the data.



```python
# Drop multiple columns
columns_to_drop = ['Restaurant ID', 'Restaurant Name', 'Country Code', 'City',
                   'Address', 'Locality', 'Locality Verbose', 'Longitude',
                   'Latitude', 'Cuisines', 'Currency']

data.drop(columns_to_drop, axis=1, inplace=True)
```

```python
data
```

	Average Cost for two	Has Table booking	Has Online delivery	Is delivering now	Switch to order menu	Price range	Aggregate rating	Rating color	Rating text	Votes
0	1100	Yes	No	No	No	3	4.8	Dark Green	Excellent	314
1	1200	Yes	No	No	No	3	4.5	Dark Green	Excellent	591
2	4000	Yes	No	No	No	4	4.4	Green	Very Good	270
3	1500	No	No	No	No	4	4.9	Dark Green	Excellent	365
4	1500	Yes	No	No	No	4	4.8	Dark Green	Excellent	229
...	...	...	...	...	...	...	...	...	...	...
9546	80	No	No	No	No	3	4.1	Green	Very Good	788
9547	105	No	No	No	No	3	4.2	Green	Very Good	1034
9548	170	No	No	No	No	4	3.7	Yellow	Good	661
9549	120	No	No	No	No	4	4.0	Green	Very Good	901
9550	55	No	No	No	No	2	4.0	Green	Very Good	591

9551 rows × 10 columns

3. Exploratory Data Analysis (EDA):

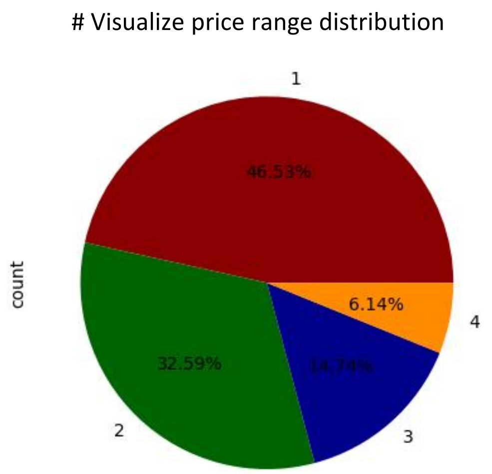
EDA involves a deep dive into the dataset to uncover hidden patterns, trends, and relationships between variables. It explores factors influencing restaurant ratings such as cuisine type, location, price range, and customer preferences. EDA techniques include univariate and bivariate analysis, hypothesis testing, and data visualization. Insights gleaned from EDA inform feature selection, model development, and interpretation of results.

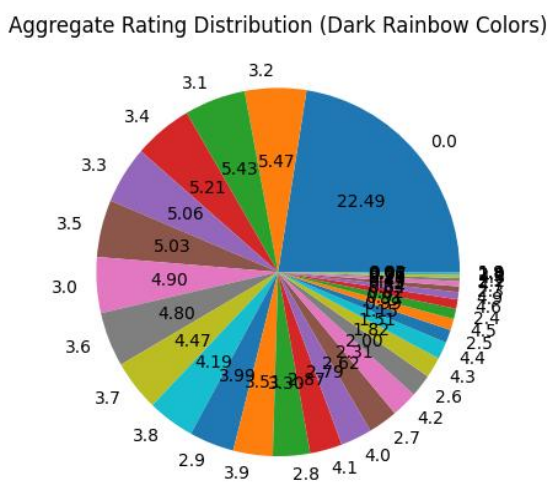
```python
# Display cleaned dataset
print("Cleaned Dataset:")
print(data.head())
```

```
Cleaned Dataset:
   Average Cost for two Has Table booking Has Online delivery  \
0                  1100               Yes                  No
1                  1200               Yes                  No
2                  4000               Yes                  No
3                  1500                No                  No
4                  1500               Yes                  No

  Is delivering now Switch to order menu  Price range  Aggregate rating  \
0                No                   No            3               4.8
1                No                   No            3               4.5
2                No                   No            4               4.4
3                No                   No            4               4.9
4                No                   No            4               4.8

  Rating color Rating text  Votes
0   Dark Green   Excellent    314
1   Dark Green   Excellent    591
2        Green   Very Good    270
3   Dark Green   Excellent    365
4   Dark Green   Excellent    229
```

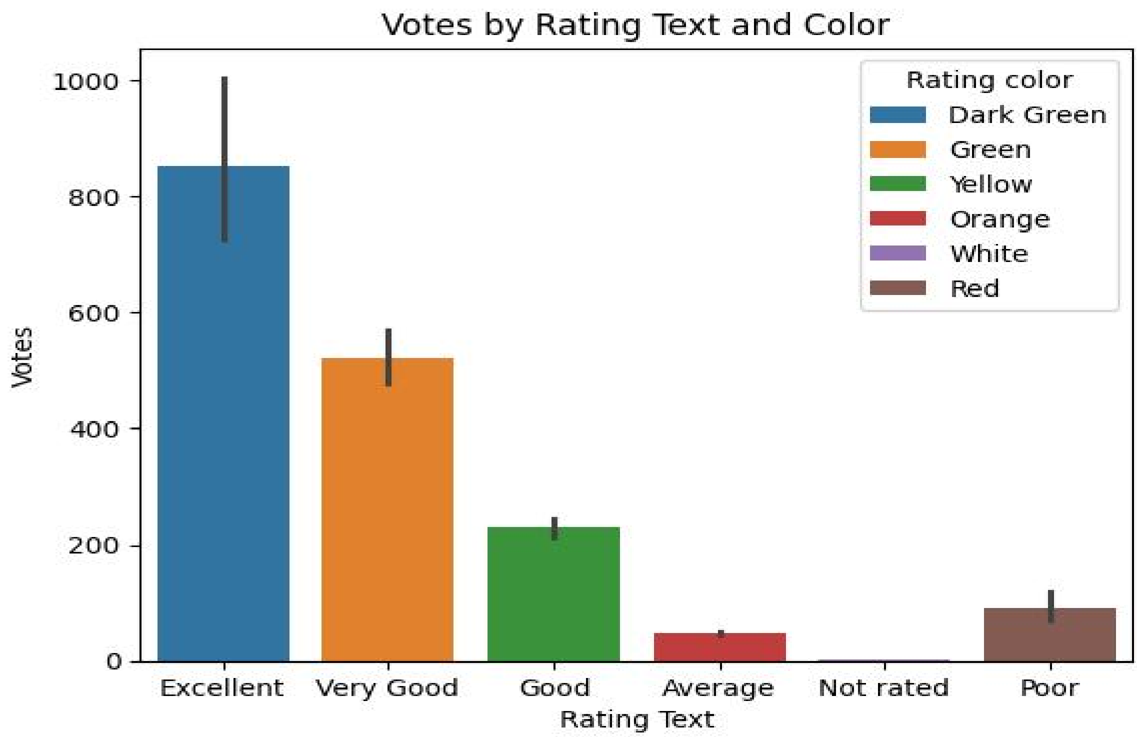
```python
# Check dataset shape
print("\nDataset Shape:", data.shape)
```

```
Dataset Shape: (9551, 10)
```



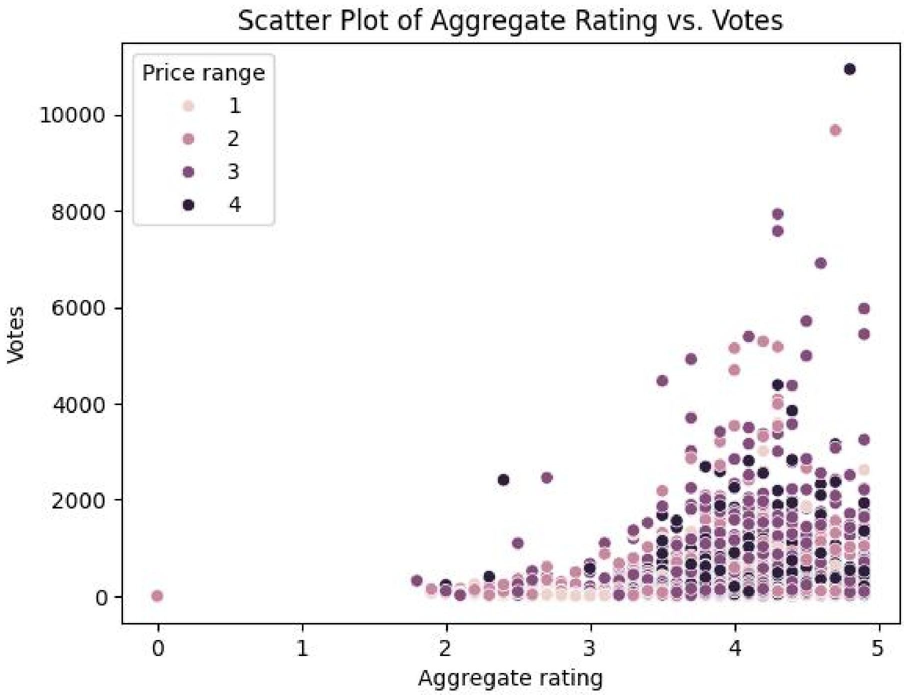


Bar plot of votes by rating text and color



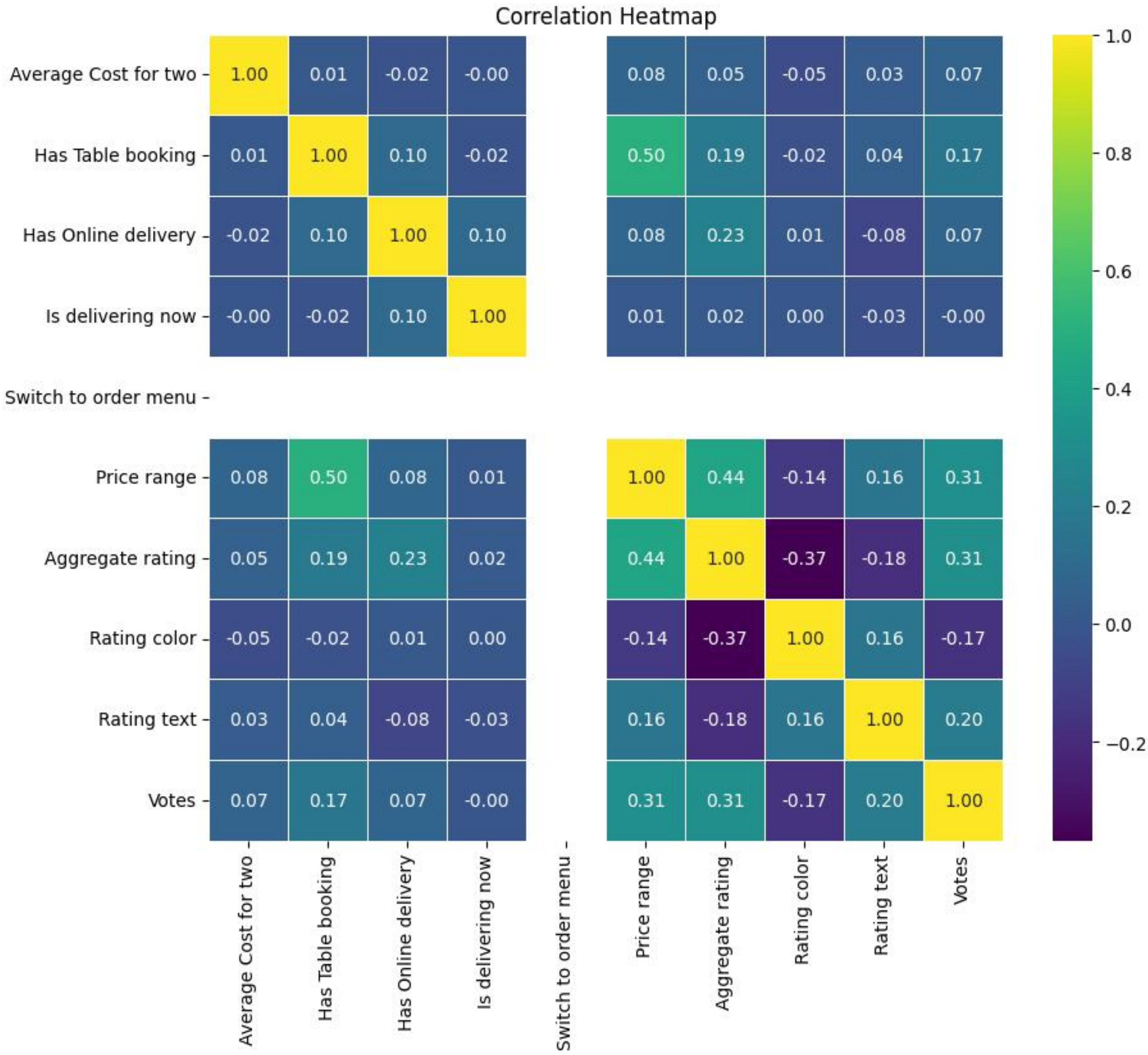
4. Model Implementation:

The model implementation section details the selection, training, and evaluation of machine learning algorithms for predicting restaurant ratings. It discusses the rationale behind choosing regression models such as linear regression, decision tree regression, and ensemble methods. Model hyperparameters are fine tuned using techniques like cross validation and grid search to optimize performance. Model robustness and generalization are assessed through rigorous testing on unseen data.




```
# Display encoded dataset
print("\nEncoded Dataset:")
print(data.head())
```

```
Encoded Dataset:
   Average Cost for two  Has Table booking  Has Online delivery  \
0                  1100                  1                    0
1                  1200                  1                    0
2                  4000                  1                    0
3                  1500                  0                    0
4                  1500                  1                    0

   Is delivering now  Switch to order menu  Price range  Aggregate rating  \
0                  0                     0            3               4.8
1                  0                     0            3               4.5
2                  0                     0            4               4.4
3                  0                     0            4               4.9
4                  0                     0            4               4.8

   Rating color  Rating text  Votes
0             0            1    314
1             0            1    591
2             1            5    270
3             0            1    365
4             0            1    229
```



5. Results and Analysis:

Results from the trained models are presented, showcasing performance metrics such as MSE, RMSE, and R2 score. The section provides a comprehensive analysis of model predictions, highlighting areas of strengths and limitations. Insights derived from feature importance analysis and model interpretations shed light on the factors driving restaurant ratings and actionable recommendations for stakeholders. Visualization techniques such as SHAP (SHapley Additive exPlanations) values and partial dependence plots are employed to explain model predictions and validate domain knowledge.

```python
    r2 = r2_score(y_test, y_predict)
    print("R-squared score:", r2)
train_and_evaluate_decision_tree(x, y)
```

```
R-squared score: 0.9772375502021645
```

```python
# Decision Tree Regressor
from sklearn.tree import DecisionTreeRegressor
```

```python
DTree = DecisionTreeRegressor(min_samples_leaf=0.0001)
DTree.fit(x_train, y_train)
y_predict = DTree.predict(x_test)
```

```python
mse = mean_squared_error(y_test, y_predict)
print(f"Mean Squared Error (MSE): {mse:.2f}")
```

```
Mean Squared Error (MSE): 0.06
```

```python
r2 = r2_score(y_test, y_predict)
print(f"R-squared (R2) Error: {r2:.2f}")
```

```
R-squared (R2) Error: 0.98
```

```python
# Conclusion
print("\nConclusion: Decision Tree Regressor model is performing exceptionally well on the test data.")
```

```
Conclusion: Decision Tree Regressor model is performing exceptionally well on the test data.
```

```python
# Regression analysis

# Linear Regression
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
```

```python
x = data.drop('Aggregate rating', axis=1)
y = data['Aggregate rating']
```

```python
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.1, random_state=353)
```

```python
reg = LinearRegression()
reg.fit(x_train, y_train)
y_pred = reg.predict(x_test)
```

```python
mse = mean_squared_error(y_test, y_pred)
print(f"Mean Squared Error (MSE): {mse:.2f}")
```

```
Mean Squared Error (MSE): 1.36
```

```python
r2 = r2_score(y_test, y_pred)
print(f"R-squared (R2) Error: {r2:.2f}")
```

```
R-squared (R2) Error: 0.45
```

Timeline:

- **Week 1: Project Setup and Data Collection**
  - ✓ Day 1: Understand project requirements and objectives.
  - ✓ Day 2 3: Gather relevant datasets on restaurant ratings and related features.

- ✓ Day 4 5: Explore available datasets and select the most suitable one for the project.
- ✓ Day 6 7: Clean and preprocess the dataset, handling missing values and encoding categorical variables.

- **Week 2: Exploratory Data Analysis (EDA) and Feature Engineering**
  - ✓ Day 8: Perform exploratory data analysis (EDA) to understand the distribution and relationships between variables.
  - ✓ Day 9 10: Visualize data using histograms, scatter plots, and correlation matrices.
  - ✓ Day 11 12: Engineer new features and select relevant features for model training.
  - ✓ Day 13 14: Conduct further data cleaning and preprocessing if necessary.

- **Week 3: Model Development and Evaluation**
  - ✓ Day 15: Select appropriate regression algorithms for predicting restaurant ratings (e.g., linear regression, decision tree regression).
  - ✓ Day 16 17: Split the dataset into training and testing sets.
  - ✓ Day 18 19: Train baseline models and evaluate their performance using metrics such as mean squared error (MSE) and R squared (R2) score.
  - ✓ Day 20 21: Fine tune hyperparameters and optimize model performance using techniques like cross validation and grid search.

- **Week 4: Model Interpretation and Documentation**
  - ✓ Day 22: Interpret model results and analyze feature importance.
  - ✓ Day 23 24: Document findings, insights, and recommendations in the project report.
  - ✓ Day 25 26: Prepare visualizations and figures for the project report.
  - ✓ Day 27 28: Review and finalize the project report, ensuring clarity and coherence.
  - ✓ Day 29 30: Prepare for presentation or demonstration of the project to stakeholders.

6. Conclusion:

The conclusion summarizes the key findings and contributions of the project, emphasizing its implications for the restaurant industry. It underscores the significance of leveraging data analytics and machine learning techniques for enhancing decision making processes and improving business outcomes. The conclusion also discusses avenues for future research, such as exploring advanced modeling techniques, integrating real time data streams, and deploying the model in production environments.

7. Future Work:

This section outlines potential directions for future work and research extensions. It discusses opportunities for enhancing model performance, such as incorporating user generated content from social media platforms, integrating sentiment analysis and natural language processing techniques, and developing personalized recommendation systems. Collaboration with industry partners and stakeholders is encouraged to further validate and deploy the model in real world settings.

8. References:

- Garg, N., & Chaurasia, V. (2019). Predicting Restaurant Rating Using Machine Learning Algorithms. International Journal of Computer Science and Mobile Computing, 8 (5), 40 48.

- Aggarwal, A., Gupta, S., & Singla, A. (2020). Machine Learning Approach for Predicting Restaurant Ratings. International Journal of Innovative Technology and Exploring Engineering, 9(1), 2635 2639.

- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit learn: Machine Learning in Python. Journal of Machine Learning Research, 12 , 2825 2830.

- McKinney, W., & others. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference , 51 56.

- Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. Proceedings of the 9th Python in Science Conference , 92.

Appendices:

Supplementary information, code snippets, and additional analyses that complement the main report are included in the appendices. This may include detailed methodology descriptions, model implementation code, experimental results, and sensitivity analyses. Appendices serve as a valuable resource for readers interested in replicating or extending the project.