Practical exam for Data Scientists

Jean-Baptiste GOMEZ

Recipe Traffic Site

01

Data Presentation

In our quest to predict which recipes will be popular on our website, we've been provided with a dataset by the product manager that includes key details on 947 recipes. This dataset, structured with seven columns, contains comprehensive nutritional information such as the number of calories, carbohydrates, sugars, and proteins each recipe contains. Additionally, it classifies each recipe into one of ten categories, ranging from 'Lunch/Snacks' to 'Desserts' and 'One Dish Meals', and also notes the number of servings. Crucially, the dataset includes a 'high_traffic' indicator, which signifies whether a recipe attracted high visitor traffic when featured on the homepage. This data will be instrumental in developing a model to predict traffic trends and, consequently, the popularity of recipes.

Data Validation

In our analysis of the recipe popularity prediction model, we've thoroughly reviewed the dataset of 947 recipes, confirming no duplicates and evaluating each column for data integrity.

We adjusted the **servings** column to maintain its numeric data type for flexibility during model development. The **high_traffic** column was converted from an object type to boolean, representing "High" and "Low" traffic based on the presence or absence of values. In the **category** column, we reclassified an inconsistent "Chicken Breast" entry under "Chicken" to align with predefined categories like 'Lunch/Snacks', 'Beverages', and others.

To address the 52 missing values in the nutritional content columns, which constitute 18.21% of our data, we filled these gaps using category and serving size-specific averages, ensuring data completeness without significant loss.

With these corrections, our dataset is now ready for exploratory analysis and subsequent descriptive statistics computation, setting the stage for accurate predictive modeling of recipe popularity.

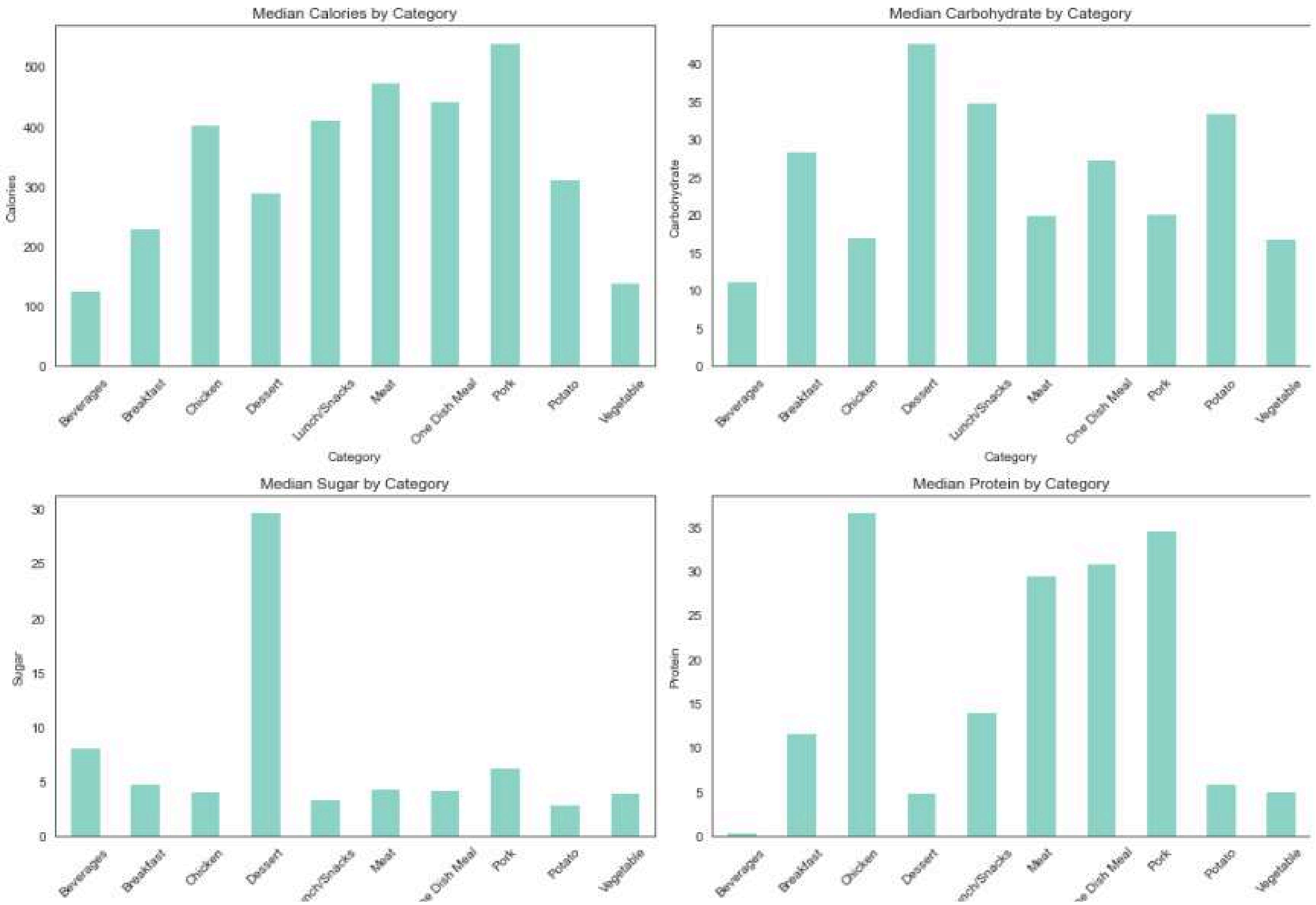
Exploratory Analysis

03

In our exploratory data analysis, the descriptive statistics have uncovered significant standard deviations in columns related to calories and nutritional components. These deviations, along with quantile-based observations of outliers, contribute to a right-skewed distribution across these metrics, suggesting that the data does not follow a normal distribution.

Exploratory Analysis

Our analysis reveals that the medians for calories and nutritional components are not symmetrical and vary depending on the type of food and beverages. We will next examine the relationship between servings and high traffic status, observing that recipes from specific categories, particularly "Vegetable," "Potato," and "Pork," tend to generate higher traffic, whereas the "Beverages" category has a minimal impact.



To further delve into this pattern, we plan to utilize binary classification algorithms from supervised machine learning to predict high traffic status. We will initially implement Logistic Regression and consider other models like Decision Tree, Random Forest, and Support Vector Machines for subsequent comparison.

Before deploying Logistic Regression, we must address outliers in our data by calculating the Interquartile Range (IQR) to manage them effectively and strengthen our model.

High values in our data require attention. Directly removing outliers causes significant data loss, so we avoid this approach. After testing transformations like Logarithmic, Square Root, and Yeo-Johnson, the Yeo-Johnson method proved most effective, especially since some columns contain zero values, which complicates other methods.

Model Development

For model development, we removed the 'recipe' ID as it is irrelevant for analysis and classification. We selected calories, carbohydrate, sugar, protein, servings, and category as features, with high_traffic as the target variable. The categorical 'category' variable was converted to a numeric feature through encoding.

Our modeling process will proceed with these steps:
1. Data Splitting: Separate the dataset into features (X) and the target variable (y), focusing on "high_traffic".
2. Train-Test Split: Use the `train_test_split` function from the scikit-learn library to divide the data into training and testing sets.
3. Model Development: Create a baseline model and compare it with additional models.
4. Model Training and Prediction: Train the chosen model on the training data and make predictions on the testing data, also checking for overfitting with the `predict` method.

These steps will guide us in creating and evaluating our baseline and comparison machine learning models.

Model Evaluation

Logistic regression	**Decision Tree**	**Random Forest**	**SVM**
[[205 91] [89 372]]	[[295 1] [2 459]]	[[294 2] [1 460]]	[[68 228] [4 457]]
Precision : 0.8	Precision : 0.72	Precision : 0.73	Precision : 0.63
Accuracy : 0.76	Accuracy : 0.67	Accuracy : 0.72	Accuracy : 0.66

Business Metrics

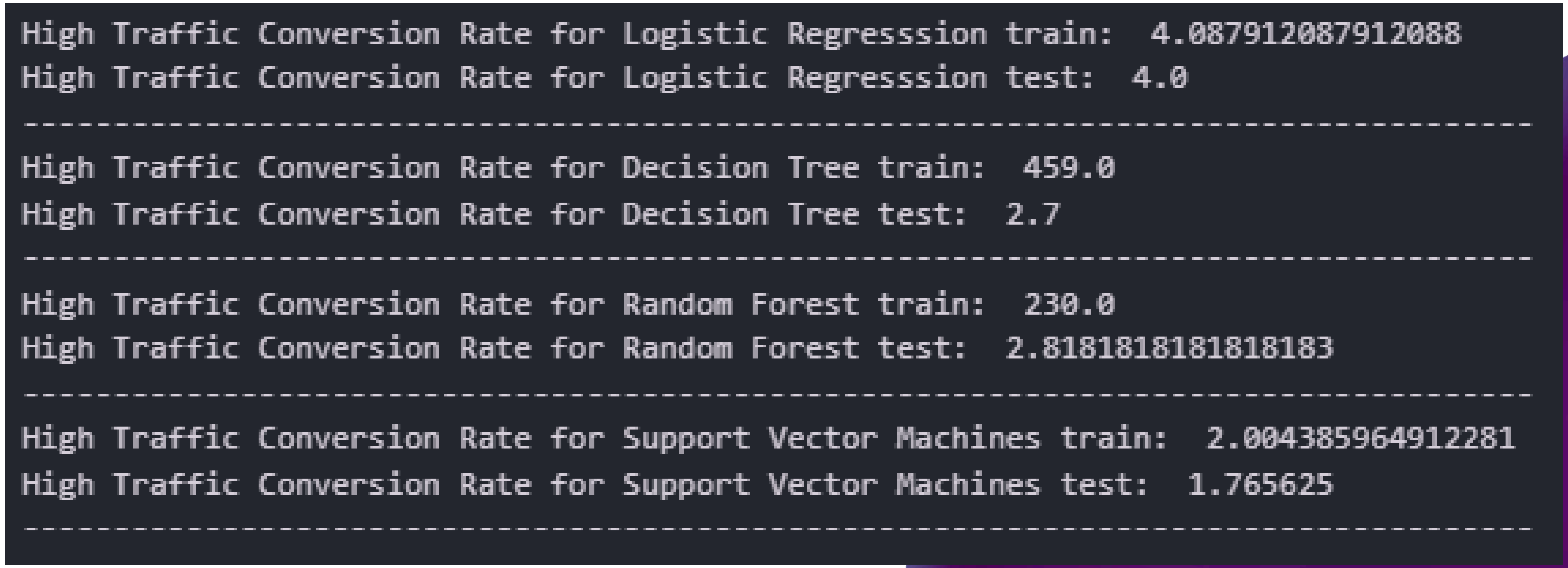
08

We have two business objectives:

1. Predict which recipes will experience high traffic.
2. Accurately forecast the "High" traffic status of recipes with at least an 80% probability.

The Logistic Regression model has successfully met both goals without the overfitting issues seen in the Random Forest and Decision Tree models. This is evidenced by its high rates of Precision, Recall, and F1 Score, all of which are 80% or higher.

04



```
High Traffic Conversion Rate for Logistic Regresssion train:  4.087912087912088
High Traffic Conversion Rate for Logistic Regresssion test:  4.0
--------------------------------------------------------------------------------
High Traffic Conversion Rate for Decision Tree train:  459.0
High Traffic Conversion Rate for Decision Tree test:  2.7
--------------------------------------------------------------------------------
High Traffic Conversion Rate for Random Forest train:  230.0
High Traffic Conversion Rate for Random Forest test:  2.8181818181818183
--------------------------------------------------------------------------------
High Traffic Conversion Rate for Support Vector Machines train:  2.004385964912281
High Traffic Conversion Rate for Support Vector Machines test:  1.765625
--------------------------------------------------------------------------------
```

Recommandation and Conclusion

To support the Product Manager in predicting high traffic for recipes, we recommend deploying the Logistic Regression Model. With an accuracy rate of approximately 81%, this model boosts confidence in effectively driving traffic across the website.

Model Deployment Steps:

1. Deployment Strategy: Outline implementation methods.
2. Data Collection: Gather diverse data to enhance learning.
3. Feature Engineering: Develop sophisticated features for better accuracy.

Key Insights:

- The Logistic Regression model proved most effective, achieving around 80% accuracy.
- We prioritized precision to avoid significant business losses from misclassification.
- We introduced a "High Traffic Conversion Rate" KPI for performance assessment.
- Categories like "Vegetable," "Potato," and "Pork" consistently drive high traffic, while "Beverages" attract less, guiding promotional strategies on the website.

These strategies and insights will guide our efforts in optimizing recipe visibility and engagement.