PRESENTATION

DataCamp Data Scientist Certification

Topic: Recipe Site Traffic

Presenter: Rezwan Islam Salvi

Project Overview and Business Goals

Objectives:

- Predict recipes that will generate **high traffic**
- Correctly predict high traffic recipes **80%** of the time

Dataset:

- A **csv** file containing 947 rows and 8 columns

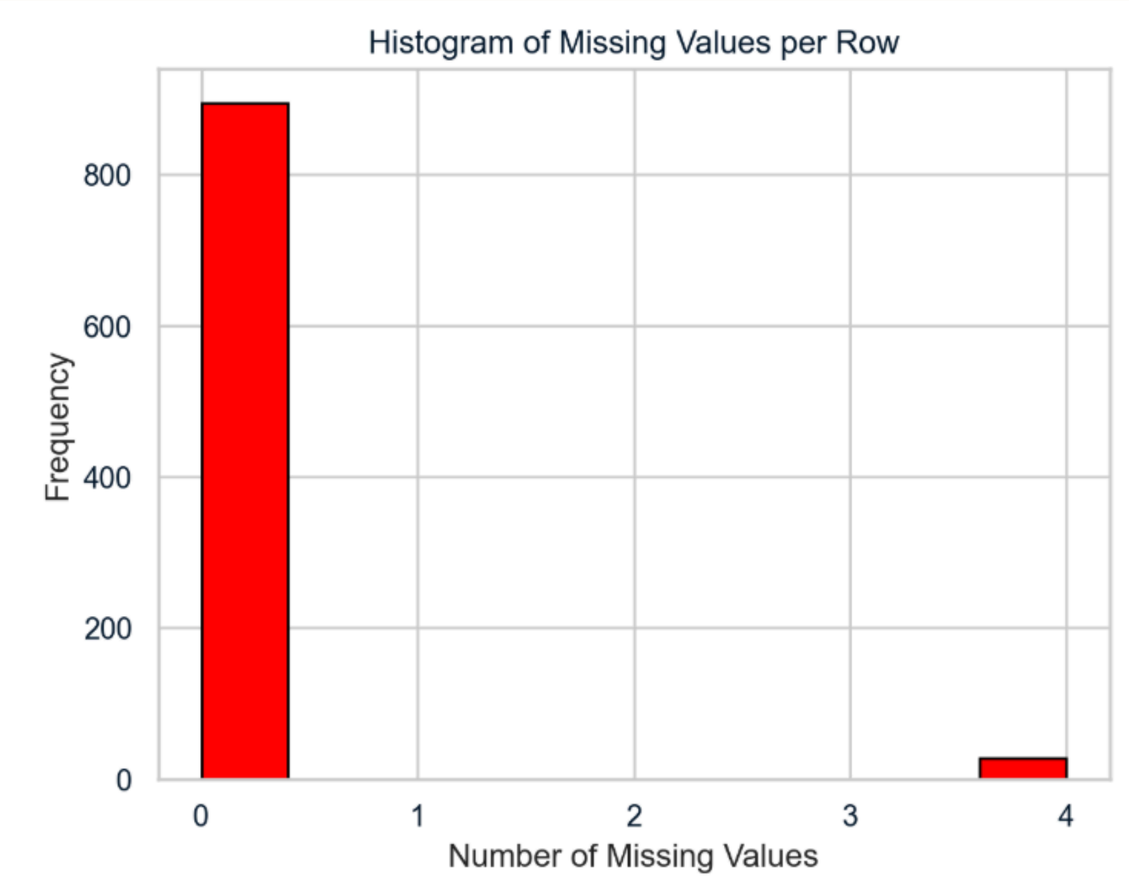
Tools used:

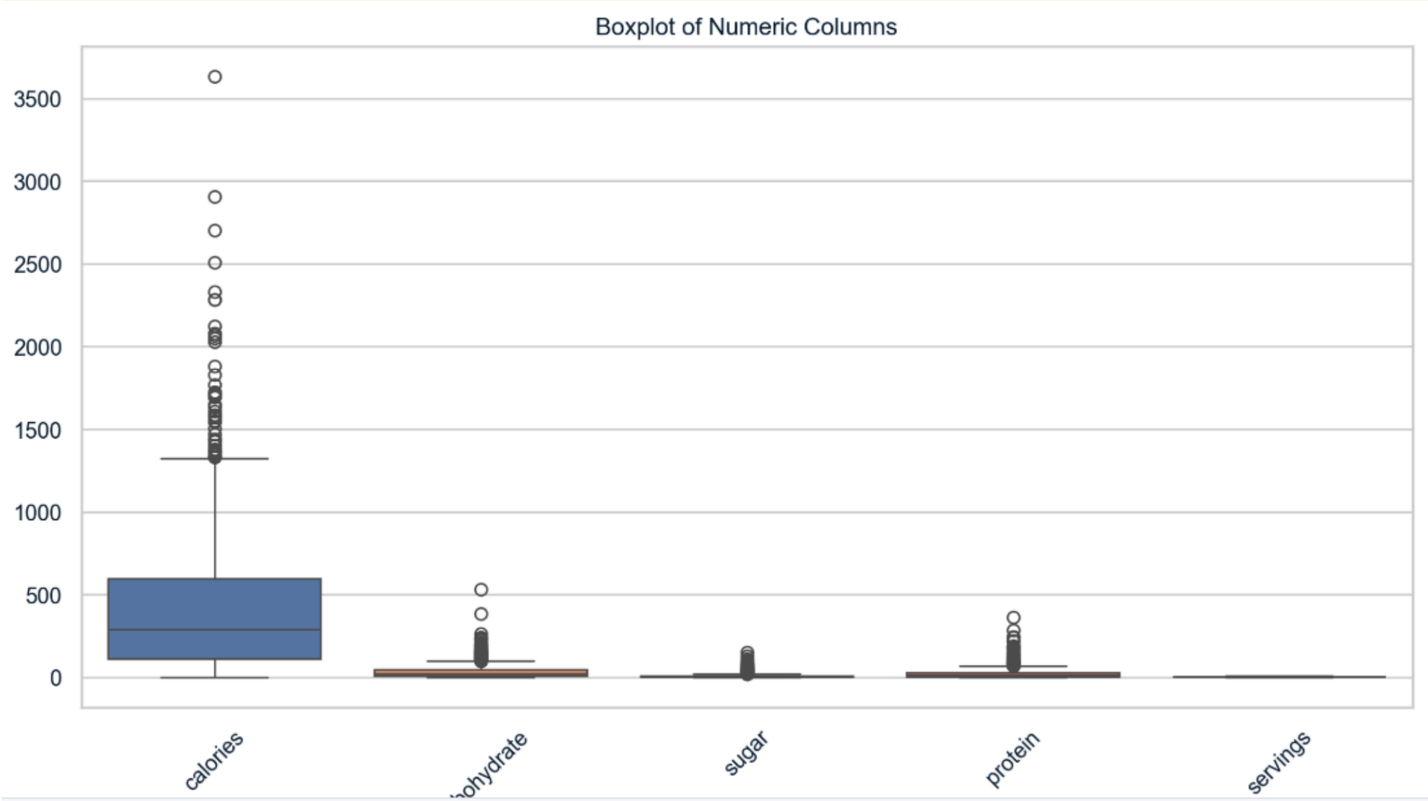
- **Python**, Data Visualization libraries, and Machine Learning models incorporated using Datalab notebook

Data Validation

- Handled **missing values** in ‘high_traffic’ column by adding ‘low’ values and renamed the column to ‘traffic’
- Converted ‘category’ and ‘traffic’ columns to **‘category’** datatype
- Replaced **non-numeric values** in ‘servings’ column with numeric equivalents
- Removed **duplicate** rows
- Removed rows with **4 missing** values
- Handled **outliers** by setting them as boundary values

Data Validation

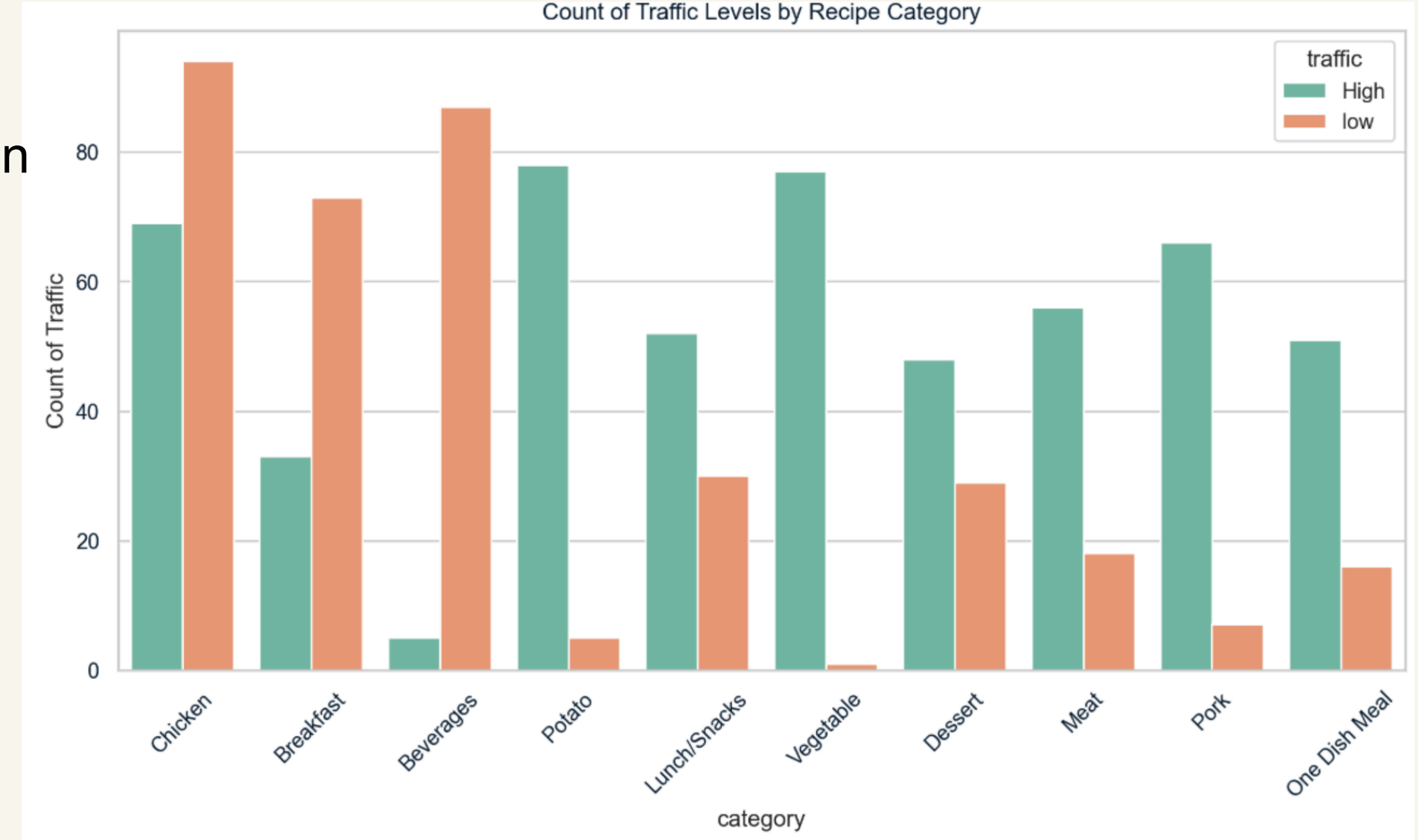




Exploratory Analysis

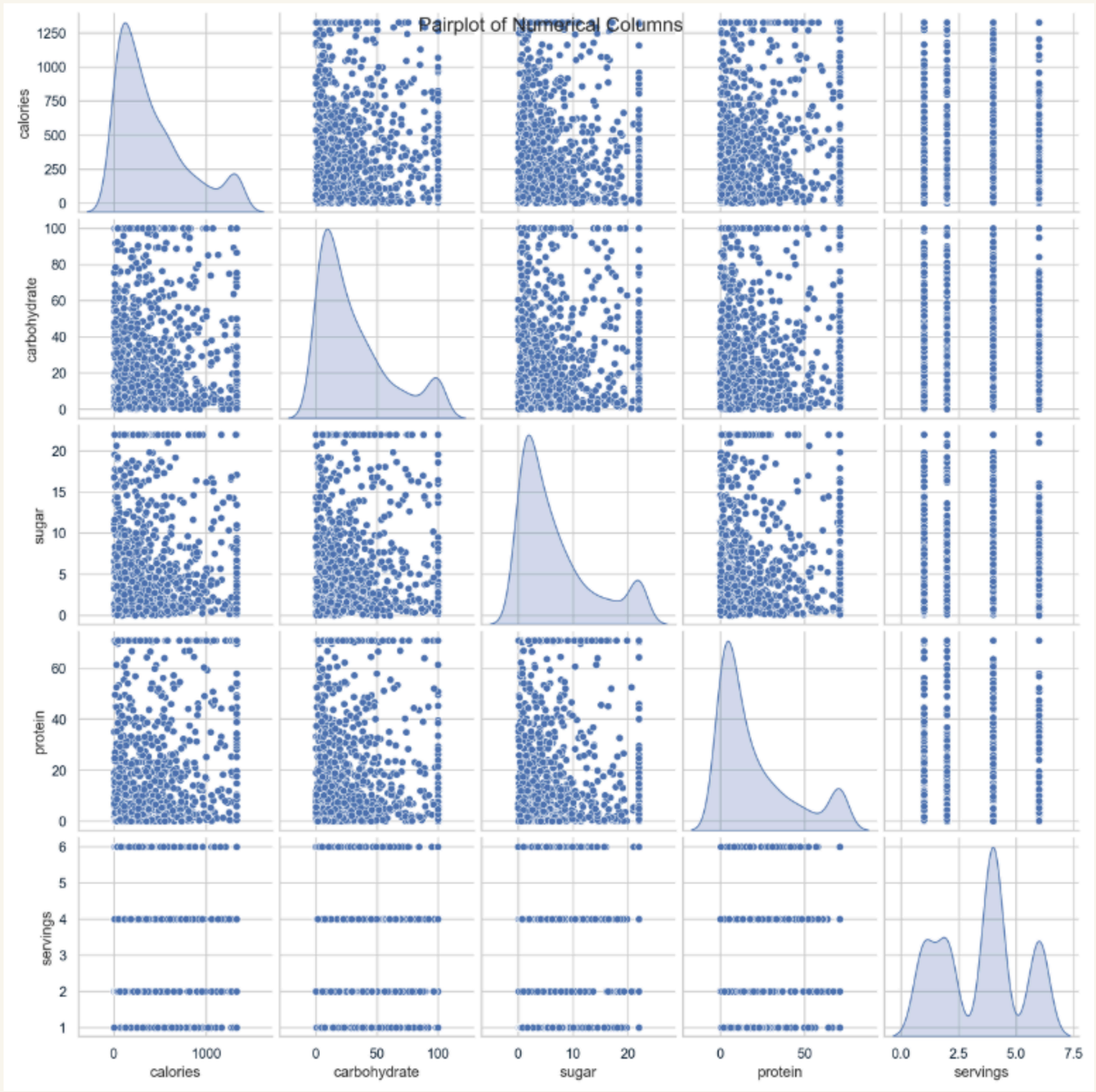
- Countplots showed '**Vegetable**' and '**Potato**' lead to **high** traffic in most occasions, and that ‘Serving’ did not have much influence

- Mean values of other numerical columns showed little influence on target



Exploratory Analysis

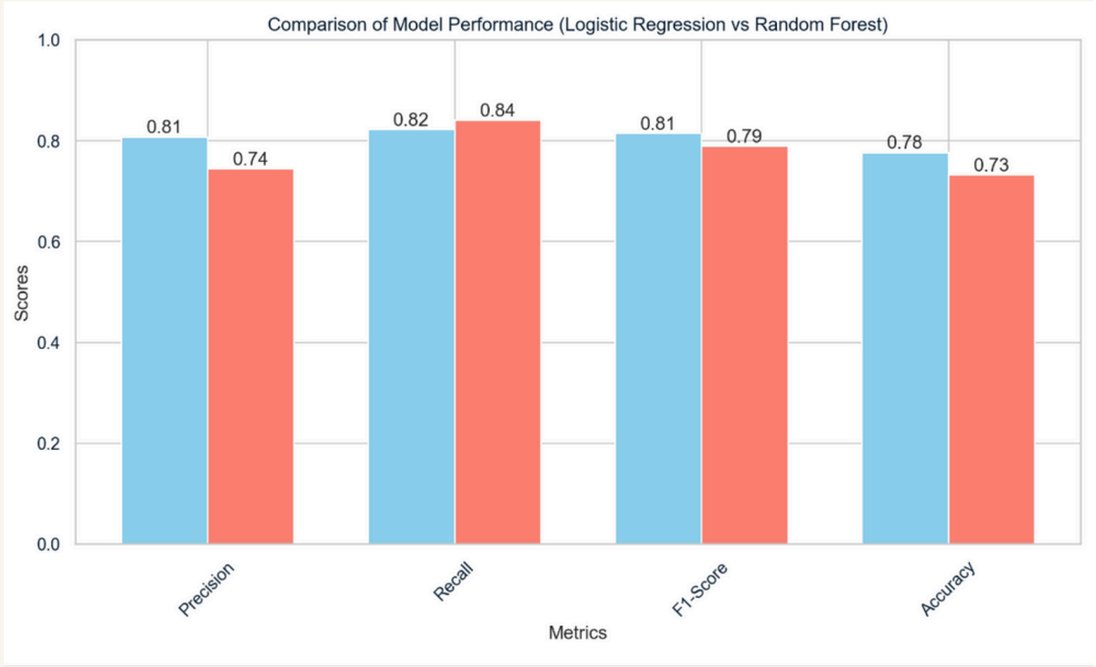
- Pairplot depicted **weak** correlation among **numeric** variables and their **right** skewness

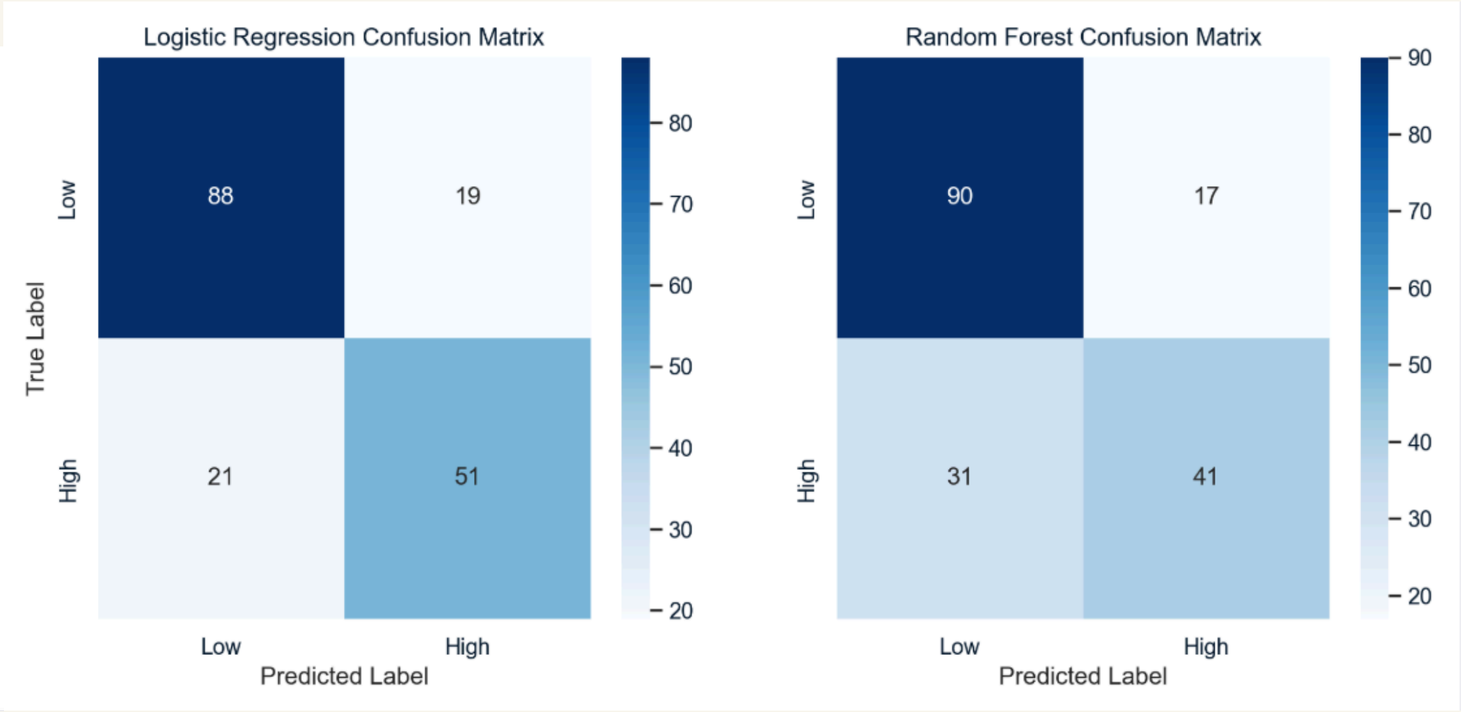


Model Development

- Performed **Feature selection** to select most significant features - category, protein, calories, and carbohydrate
- Split dataset into **80-20** ratio for training and testing
- **Standardized** the numerical columns and **One-hot encoded** the category column
- Deployed **Logistic Regression** and **Random Forest** models as they are efficient in classification tasks

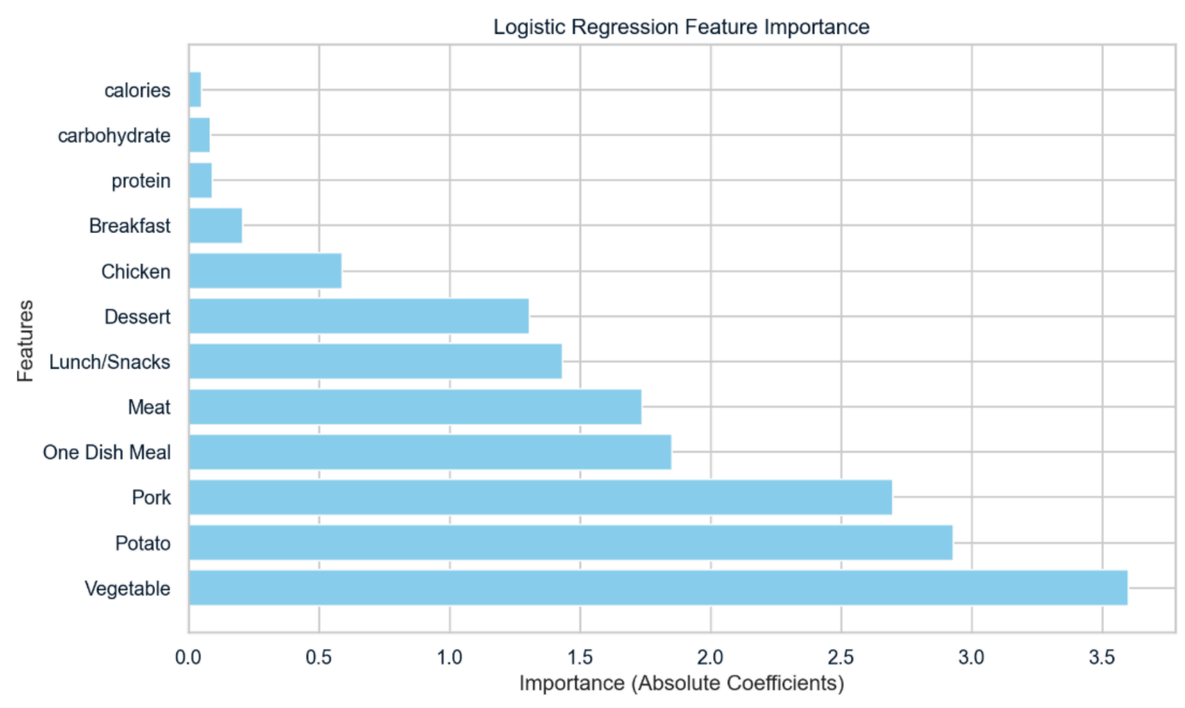
Model Evaluation and Business Metrics

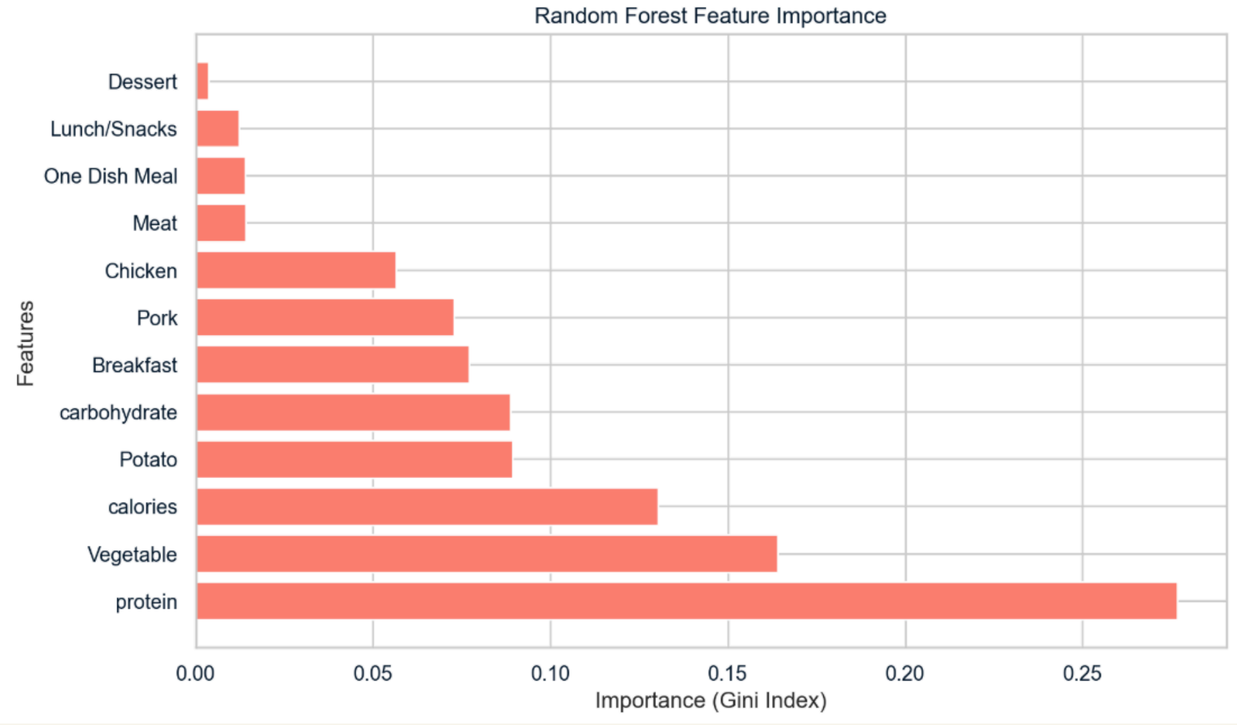




- We chose **'F1-score**' as the KPI
- **Logistic Regression** was the better model

Model Evaluation and Business Metrics





- **Vegetable**, **Potato** and **Protein** were the most important features

Recommendations

- Focus on **Key Features**
- **Real-Time** Traffic Insights
- Targeted **Marketing** and Customer Engagement
- Improve **Operational** Efficiency
- Address **Class Imbalance** in Data

THANK YOU