Data Scientist Professional Practical Exam DS601P

TASTY BYTES RECIPE TRAFFIC FORECASTING

WRITTEN BY BRANDON MCKIMMY

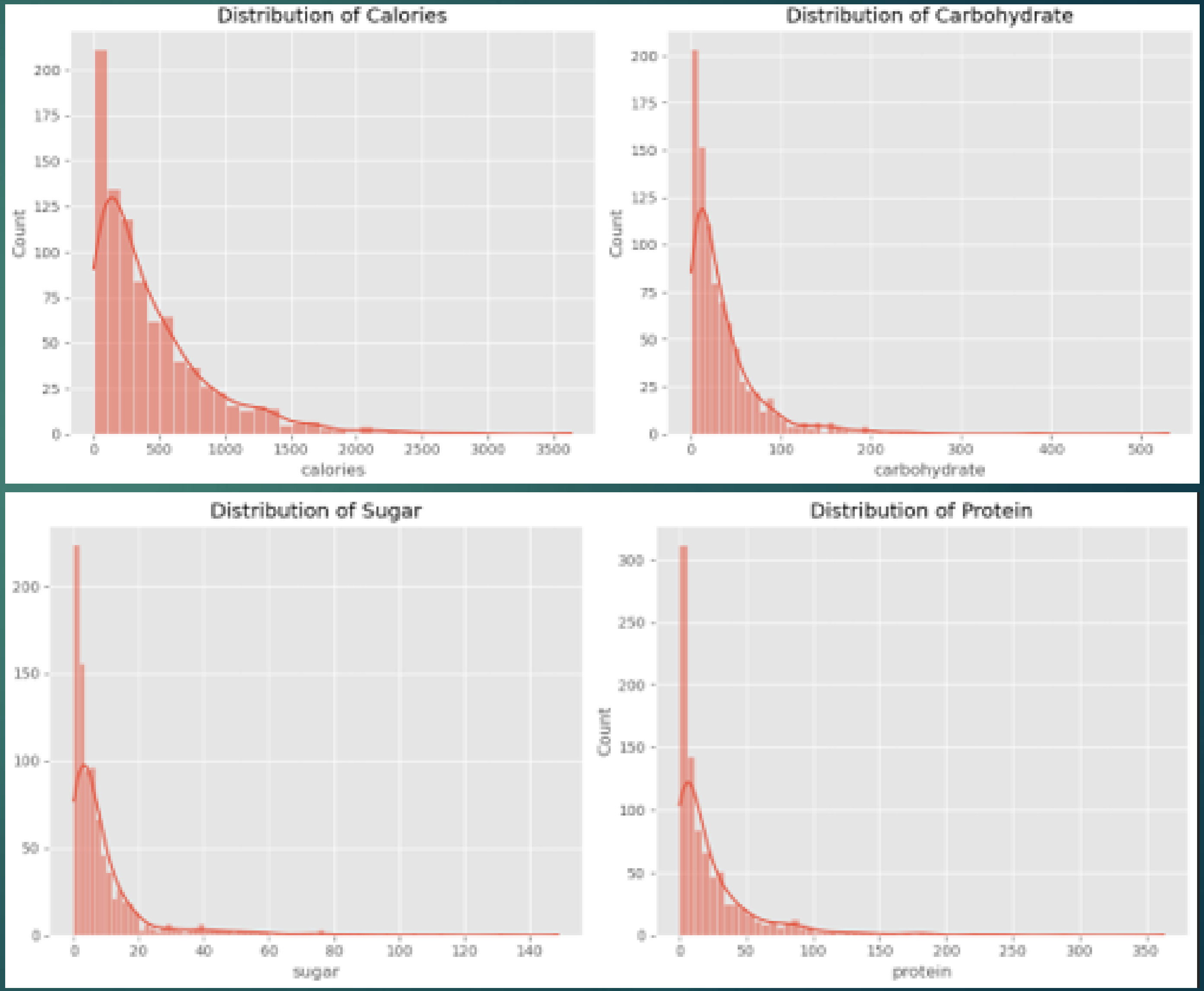
DATE: JANUARY 28TH, 2024

Exploratory Data Analysis

Nutritional Content Insights:

- Calories: Most recipes have lower calories, with a few high-calorie outliers, indicating standard preference ranges.
- Carbohydrates: Dominance of lower-carb recipes, with a drop in frequency at higher values, guiding content strategy.
- Sugar: Lower sugar levels are more common, possibly reflecting dietary trends or editorial choices.
- Protein: A diverse range of protein levels suggests a variety of dietary preferences and needs.

These patterns guide our strategy to align with audience trends and preferences.


Distribution of Calories

Distribution of Carbohydrate

Distribution of Sugar

Distribution of Protein

Exploratory Data Analysis (cont.)

Recipe Category Distribution:

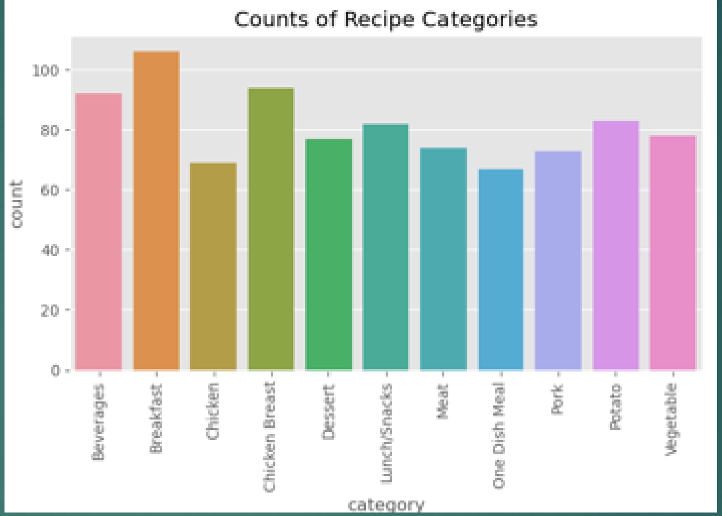
- Categories 'Breakfast', 'Chicken Breast', and 'Beverages' are most frequent.
- Suggests user engagement may be influenced by these popular categories.
- Potential for targeted content development and marketing.

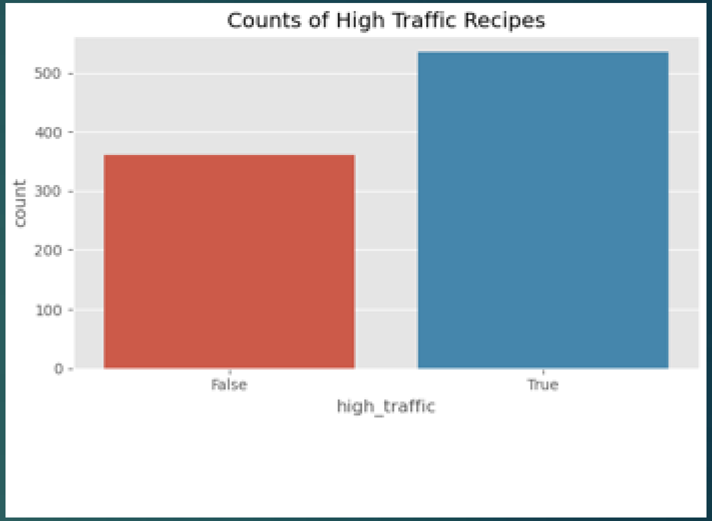
High Traffic Recipe Distribution:

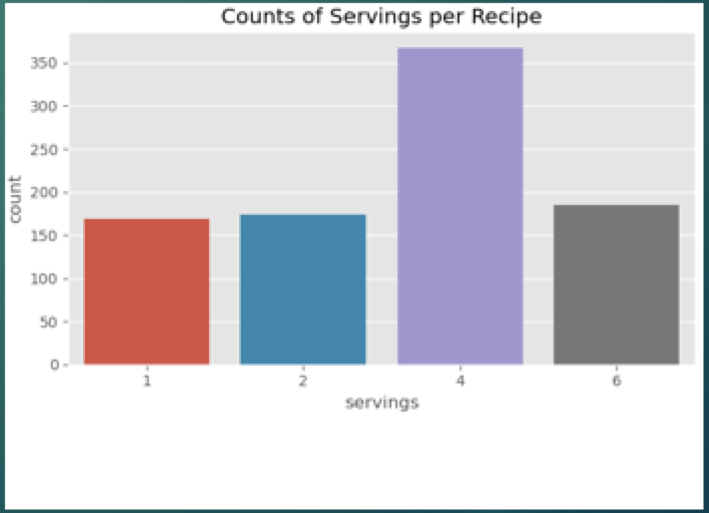
- More recipes are marked 'True' for high traffic, indicating successful user engagement.
- Insights from this distribution can refine predictive models and content strategy.

Serving Size Preferences:

- Recipes for four servings are most common, reflecting potential user preference.
- Serving size trends can inform content development to align with audience needs.



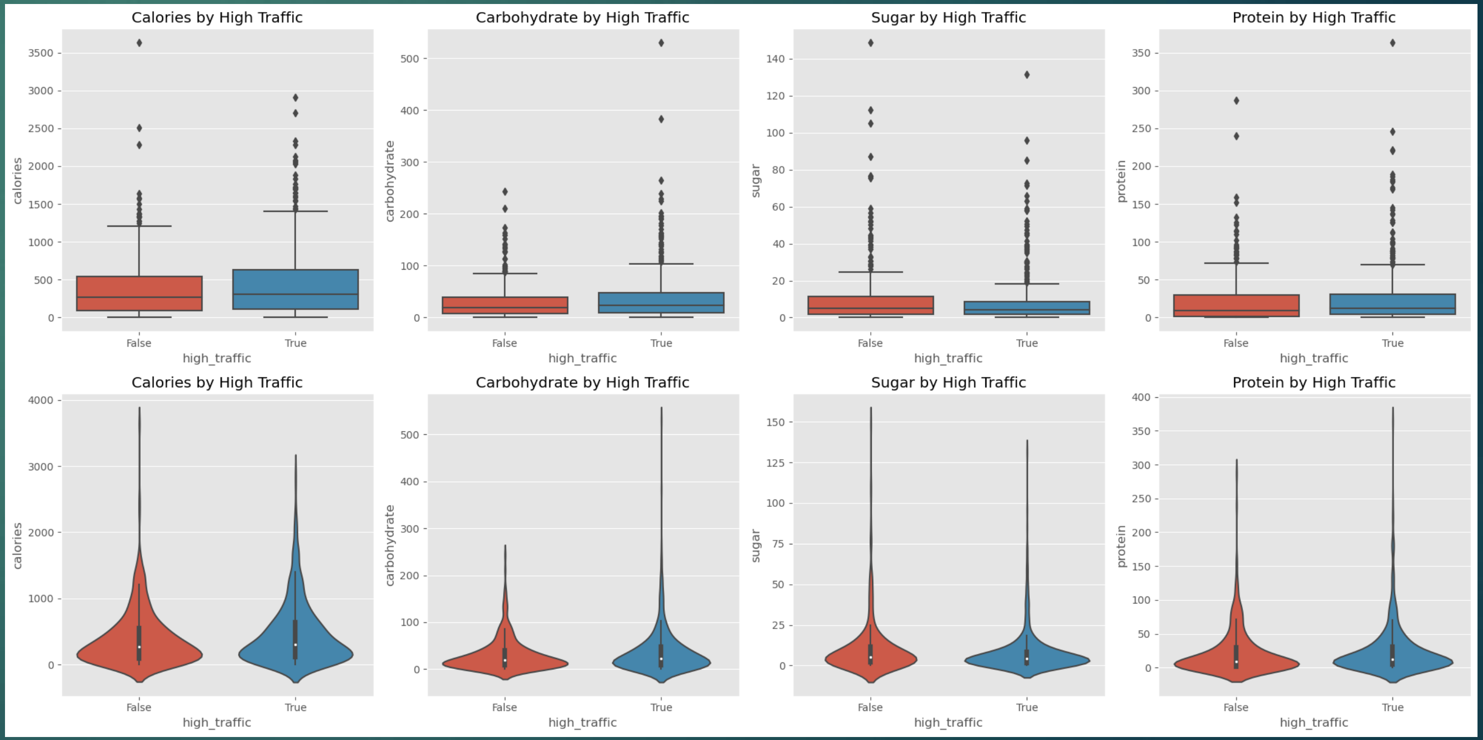




Exploratory Data Analysis (cont.)

Analytical Insights from Distribution Visualizations by High Traffic Status:

- High-traffic recipes tend to have a higher median calorie and carbohydrate content, suggesting these factors may influence user engagement.
- Sugar content shows a modest difference, hinting at its lesser impact on popularity.
- Protein levels do not significantly differ with traffic status, indicating the need for further analysis.
- The broader distribution of calories and carbohydrates in popular recipes revealed by violin plots suggests their potential as popularity indicators.
- Strategic modeling should consider these insights, with calorie and carbohydrate content as key features, while sugar and protein's roles require deeper exploration.



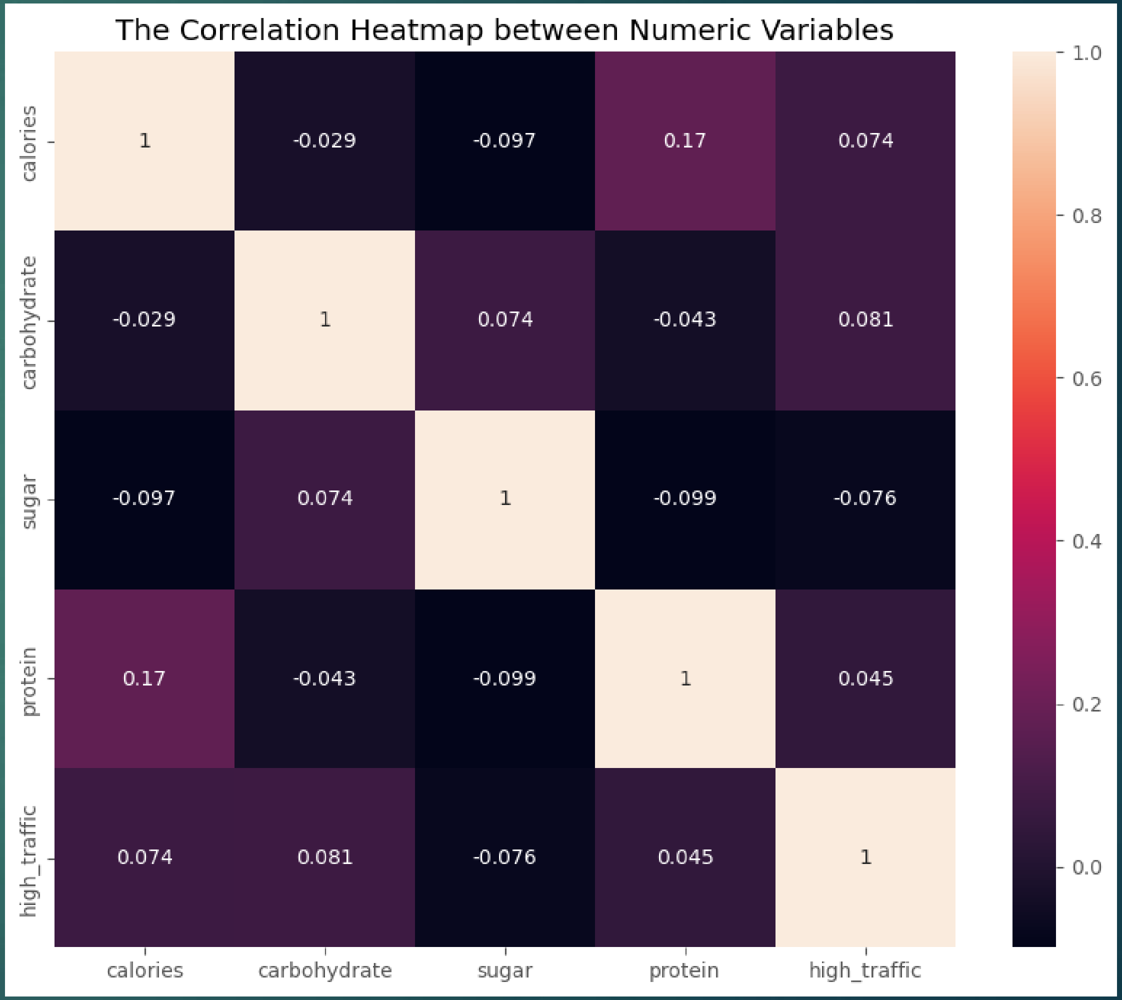
Exploratory Data Analysis (cont.)

Correlation Insights with High Traffic

- **Calories**: Slight positive correlation (0.074), but not a strong traffic predictor.
- **Carbohydrate**: Weak positive correlation (0.081), marginally influencing traffic.
- **Sugar**: Weak negative correlation (-0.076), suggesting less sugar may slightly favor traffic.
- **Protein**: Very weak positive correlation (0.045), unlikely to impact traffic alone.

Exploratory Data Analysis Summary & Model Feature Selection:

- Weak correlations suggest no single nutritional factor, like 'calories' or 'sugar', dominantly predicts high traffic.
- Traffic seems influenced by multiple recipe attributes, requiring a holistic modeling approach.
- For model fitting, I'll focus on features like 'calories', 'carbohydrate', and popular categories ('Pork', 'Potato', 'Vegetable').
- These insights will guide the selection of predictors to enhance our chances of meeting the 80% accuracy target for high traffic recipes.
- This summary retains the core findings and next steps, ensuring the presentation remains focused on both the results of the EDA and how they inform the model fitting process.



The Correlation Heatmap between Numeric Variables

	calories	carbohydrate	sugar	protein	high_traffic
calories	1	-0.029	-0.097	0.17	0.074
carbohydrate	-0.029	1	0.074	-0.043	0.081
sugar	-0.097	0.074	1	-0.099	-0.076
protein	0.17	-0.043	-0.099	1	0.045
high_traffic	0.074	0.081	-0.076	0.045	1

Optimizing Predictive Models: Strategies and Performance

- Our model fitting strategy is anchored in using Logistic Regression and Linear Discriminant Analysis (LDA) as our foundational algorithms. Logistic Regression is chosen for its simplicity and effectiveness in binary classification problems, like predicting high traffic for recipes. It serves as our baseline to compare more complex models. Meanwhile, LDA is employed to take advantage of its ability to model the difference between high and low traffic recipes by finding a linear combination of features that characterizes or separates two classes.
- To refine these models, I implement GridSearchCV and RandomizedSearchCV, which are systematic methods for tuning hyperparameters. GridSearchCV exhaustively searches through a specified parameter grid, ensuring that I test the model across all combinations of the parameter space, thus identifying the most optimal settings. In contrast, RandomizedSearchCV randomly samples a given number of parameter settings from the specified distributions, which offers a faster, stochastic approach to parameter tuning.
- By combining these models and tuning methods, I aim to achieve a balance between predictive accuracy and computational efficiency, ultimately moving towards a model that meets our business needs of accurately predicting high-traffic recipes."

Model Performance, Hyperparameter Tuning Insights, and Evaluation

Model Performance Overview

- **Logistic Regression**: Provided baseline accuracy, underperformed against the 80% target.
- **GridSearchCV (LR)**: Improved Logistic Regression, yet didn't hit the target.
- **RandomizedSearchCV (LR)**: Comparable to GridSearchCV, still below target accuracy.

Advanced Model Insights

- **Linear Discriminant Analysis**: Slightly better than Logistic Regression but didn't reach 80%.
- **GridSearchCV (LDA)**: Best performance among all models, but still below the target.
- **RandomizedSearchCV (LDA)**: Lower accuracy, indicating GridSearchCV's superior tuning.

Key Takeaways:

- Hyperparameter tuning showed performance gains.
- No model achieved the desired 80% accuracy.
- Further exploration with advanced techniques and richer feature engineering recommended.

Key Findings

- **Baseline Accuracy**: Logistic Regression models served as a benchmark, with initial accuracy around 64.80%. Hyperparameter tuning provided marginal improvements.
- **Linear Discriminant Analysis (LDA) Performance**: The basic LDA model scored slightly lower than Logistic Regression. However, GridSearchCV tuning enhanced LDA's accuracy significantly to 68.30%, while RandomizedSearchCV tuning reduced it to 59.92%.
- **Hyperparameter Impact**: Tuning through GridSearchCV and RandomizedSearchCV demonstrated variable performance enhancements, affirming the value of methodical hyperparameter optimization.
- **Model Suitability & Limitations**: Despite robustness, Logistic Regression models reached a performance ceiling. LDA showed potential for higher accuracy, especially with GridSearchCV.
- **Towards 80% Accuracy**: No model achieved the target 80% accuracy, suggesting the need for more complex modeling strategies or improved feature engineering.
- **Actionable Insight**: The analysis emphasizes the need for comprehensive model selection and tuning to enhance predictive capabilities, with further model refinement as a pathway to better performance.

Business Focus: Enhancing Content Strategy with Predictive Analytics

- **Enhancing Content Strategy with Predictive Analytics**
  - **Objective:** Employ data analytics to identify high-traffic recipe trends for Tasty Bytes, driving strategic content development and user engagement.
  - **Method:** Developed predictive models including Logistic Regression and LDA, enhanced with hyperparameter tuning techniques such as GridSearchCV and RandomizedSearchCV.
  - **Impact:** These models form a data-driven foundation for anticipating user preferences, integral to refining content strategies and boosting site traffic.
  - **Direct Impact**: Developed predictive models (Logistic Regression, LDA) with advanced hyperparameter tuning to classify potential high-traffic recipes.
- **Future Recommendations**:
  - **Data Enrichment**: Expand data collection to refine predictions, considering user feedback and broader culinary trends.
  - **Advanced Methods**: Pursue sophisticated algorithms for improved accuracy beyond the current 80% goal.
  - **Customized Strategy**: Adapt content to user preferences for increased engagement.
  - **Iterative Refinement**: Continuously update models to keep pace with changing tastes and preferences.
  - **Engagement Integration**: Merge model insights with engagement metrics for a holistic content strategy.
- By adopting these strategies, Tasty Bytes aims to sharpen its content focus and align offerings with audience demand.

Business Metrics

- **Key Performance Indicator (KPI) Analysis and Model Performance**
- **KPI Focus**: Accuracy in predicting high-traffic recipes, with a target rate of 80% to inform content strategy.
- **Logistic Regression (LR) Insights**:
  - Baseline LR achieved ~64.8% accuracy.
  - GridSearchCV's fine-tuning slightly increased LR accuracy.
  - RandomizedSearchCV showed comparable results to GridSearchCV.
- **Linear Discriminant Analysis (LDA) Insights**:
  - Baseline LDA recorded ~62.6% accuracy.
  - GridSearchCV notably improved LDA's accuracy to ~68.3%.
  - RandomizedSearchCV tuning resulted in a reduced accuracy.
- **Strategic Takeaways**:
  - The gap between current model performance and the KPI indicates room for advanced analytical methods.
  - Continuous optimization and advanced modeling techniques are recommended for achieving the KPI.

Project Goals & Model Performance Summary

- **Primary Goal**: Develop a model to predict high-traffic recipes, aligning with business strategy to boost user engagement and traffic.
- **Approach**: Utilized Logistic Regression and LDA models, enhanced with GridSearchCV and RandomizedSearchCV for optimization.
- **Outcomes**: Achieved modest accuracies around 65-68.3%, falling short of the 80% target.
- **Key Challenge**: Improving accuracy to meet the 80% benchmark for reliable predictions.
- **Recommendations**:
  - Consider advanced modeling and feature engineering.
  - Integrate additional data points for richer insights.
  - Maintain ongoing model refinement to adapt to user trends.