Practical Exam - Recipe Site Traffic

BY DANIIL MOROZKOV

Business goals

Improve existing predictive models

Imporevement of existing predictive models can help to make better decisions, reduce costs, and improve operational efficiency.

Improve web-site traffic

By deciding the most important factors which contirbute to high traffic we will be able to attract more new users on the website which will result in increased profits

Develop business metric

The goal is to create feasible business metric which will be used to monitor current performance of the web-site and will help to understand what steps are leading to better results

Project Pipeline

Data Validation

Upload, clean and preprocess the data

Exploratory Analysis

Find useful insights from visual data analysis

Model Development

Define the problem type, select the baseline and alternative models

Model Evaluation

Describe the performance of the two models based on an appropriate metric

Business Metrics development

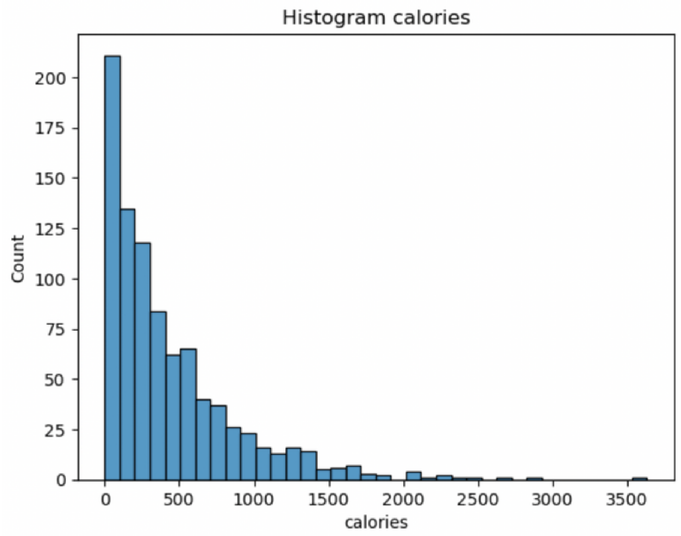
Develop a business metric based on the model designed in previous steps

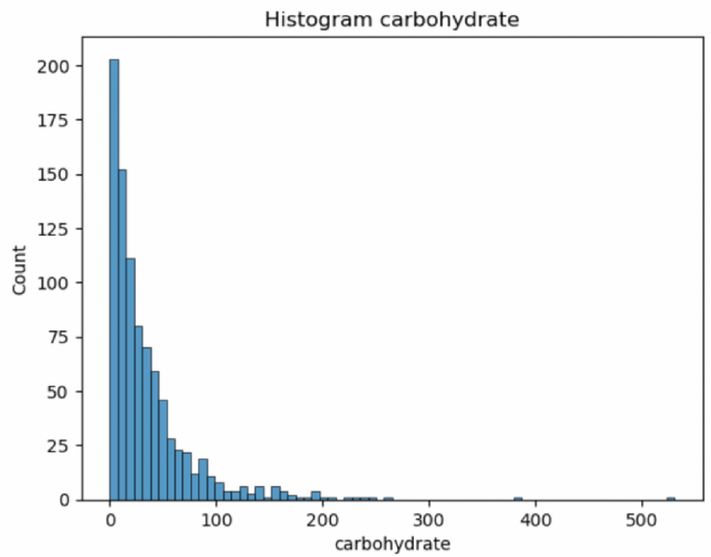
DATA PREPROCESSING SUMMARY

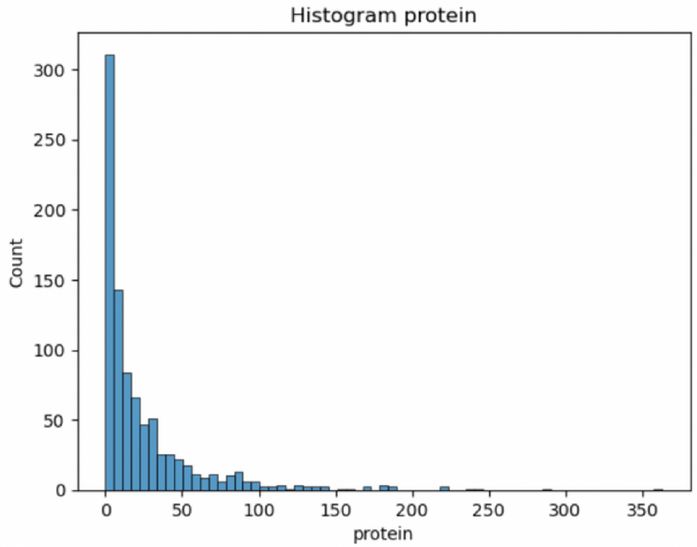
Backlog of changes

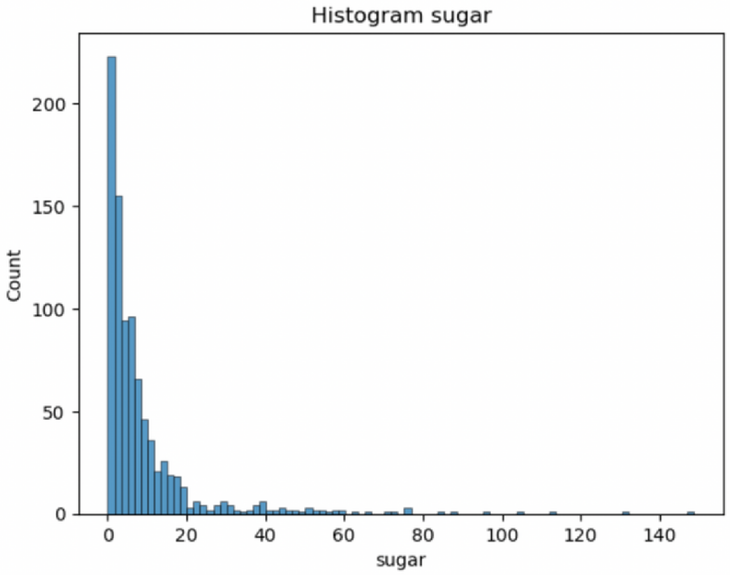
- **calories** - NA rows, extreme outliers (Q3 + 3 * IQR) were capped, scaled using MinMax
- **carbohydrate** - NA rows dropped, extreme outliers (Q3 + 3 * IQR) were capped, scaled using MinMax
- **sugar** - NA rows dropped, extreme outliers (Q3 + 3 * IQR) were capped, scaled using MinMax
- **protein** - NA rows dropped, extreme outliers (Q3 + 3 * IQR) were capped, scaled using MinMax
- **category** - 'Chicken Breast' values for changed to 'Chicken', then columns were one hot encoded
- **servings** - 2 values contained 'as a snack', they belonged to Lunch/Snack category so 'as a snack' was removed, the column was transformed to float 64
- **high_traffic** - NaN values were replaced with 'not high' and then changed to 1 - high, 0 - not high

EXPLORATORY DATA ANALYSIS

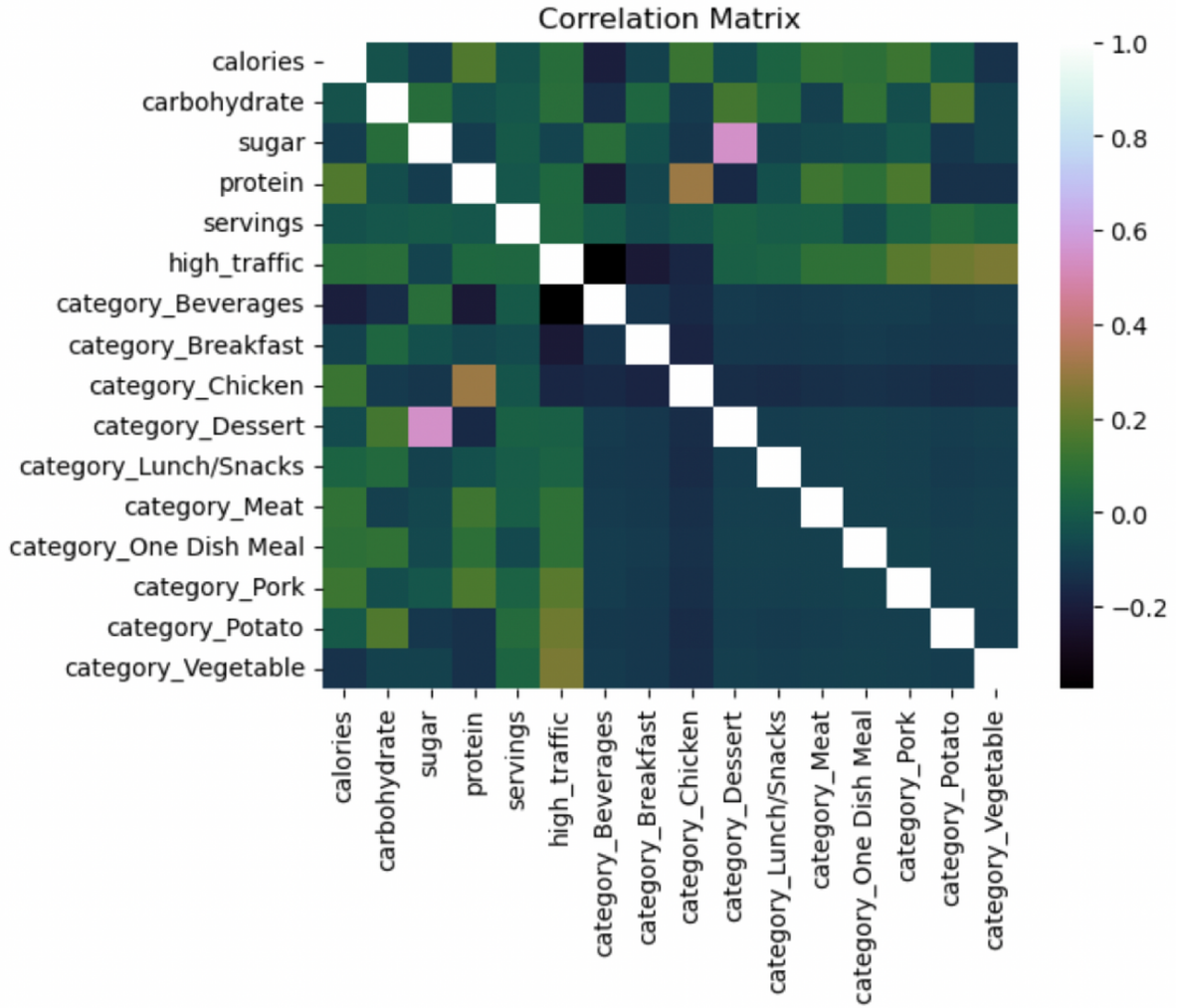


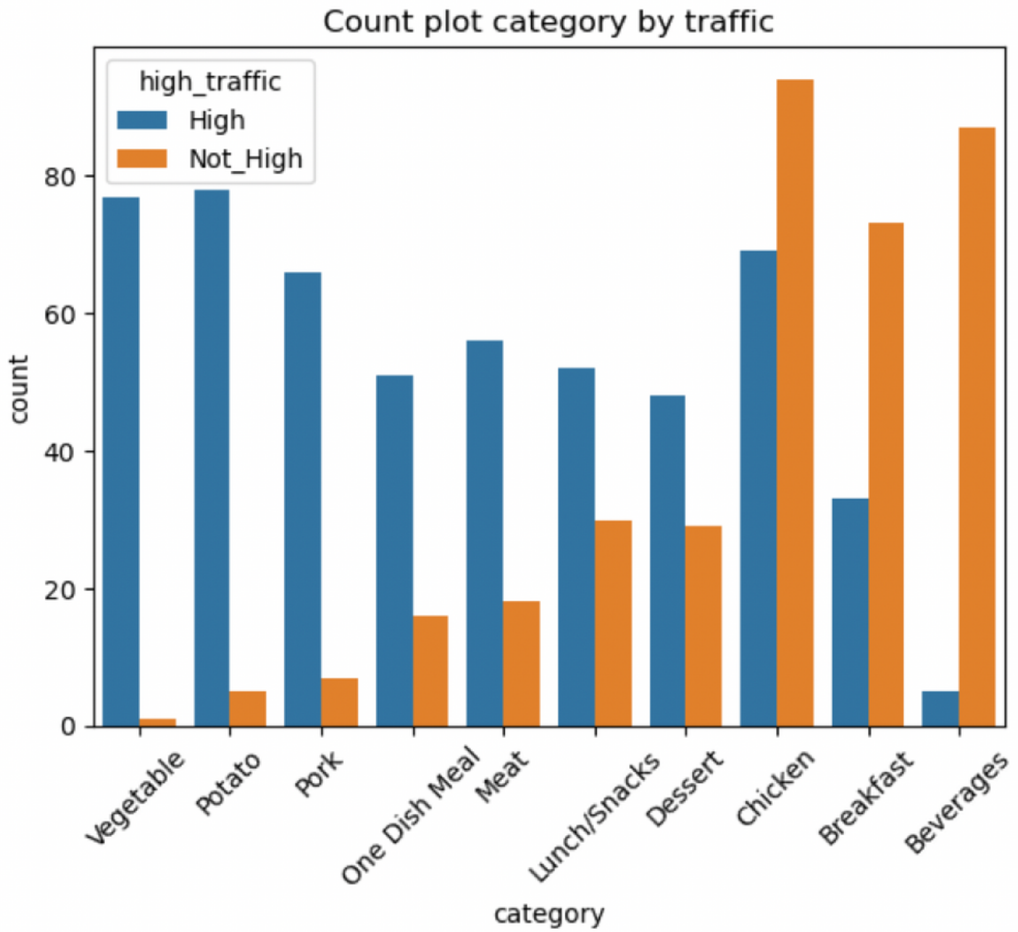






EXPLORATORY DATA ANALYSIS





MODEL DEVELOPMENT

CLASSIFICATION PROBLEM

(1 high traffic, 0 - low traffic)

Initial list of models

LogisticRegression, KNeighborsClassifier,GaussianNB, DecisionTreeClassifier,RandomForestClassifier ,LinearSVC, Perceptron, SGDClassifier.

Evaluation Metrics

Cross Validation (k=10) F1, precision and recall over 80% were the main metrics to assess the quality of the models.

MODEL EVALUATION

Logistic Regression

CV f1 score is 0.7979

CV precision is 0.797

CV recall is 0.8014

Stochastic Gradient Descent Classifier

CV f1 score is 0.7738

CV precision is 0.7616

CV recall is 0.8952

After grid search

```
Y_test Accuracy = 0.7806691449814126
Y_test ROC Area under Curve = 0.7771125555935683
Confusion matrix:
[[ 84  27]
 [ 32 126]]

Classification report:
              precision    recall  f1-score   support

           0       0.72      0.76      0.74       111
           1       0.82      0.80      0.81       158

    accuracy                           0.78       269
   macro avg       0.77      0.78      0.78       269
weighted avg       0.78      0.78      0.78       269
```

Business metric

$$\textbf{TRAFFIC RATE} = \frac{\text{Probability that traffic will be high (predictied by the developed logistic regression)}}{\text{Number of observations}}$$

CONCLUSION AND RECOMMENDATIONS

RECCOMENDATIONS

1. Show categories that have great positive correlation such as vegetable, potato recipes
2. Show less categories that have negative correlation such as beverages, breakfast
3. Show more recipes that have more servings, good on calories, protein and carbohydrate and contain less sugar. ('healthy recipes')
4. Collect data on missing values to improve the performance of the model

THANK YOU!