Recipe Site Traffic Case Study

Muhammad Atef

Business Problem

- Currently the Product Manager chooses their favorite from a selection and displays that on the home page.

- They have noticed that traffic to the rest of the website goes up by as much as 40% if they pick a popular recipe. But they don’t know how to decide if a recipe will be popular.

- “More traffic means more subscriptions so this is really important to the company. ”

Business Goals

- To increase website traffic by displaying popular recipes on the homepage.

- This leads to increase subscriptions which is essential for the business

- To use data science to predict which recipes will lead to high traffic and correctly predict high traffic recipes 80% of the time.

Data Validation

- Checked the data against the given description:

1- high_traffic has many ‘NULLs’ and ‘High’ that means not high and high:

- change “nan” to “Low”

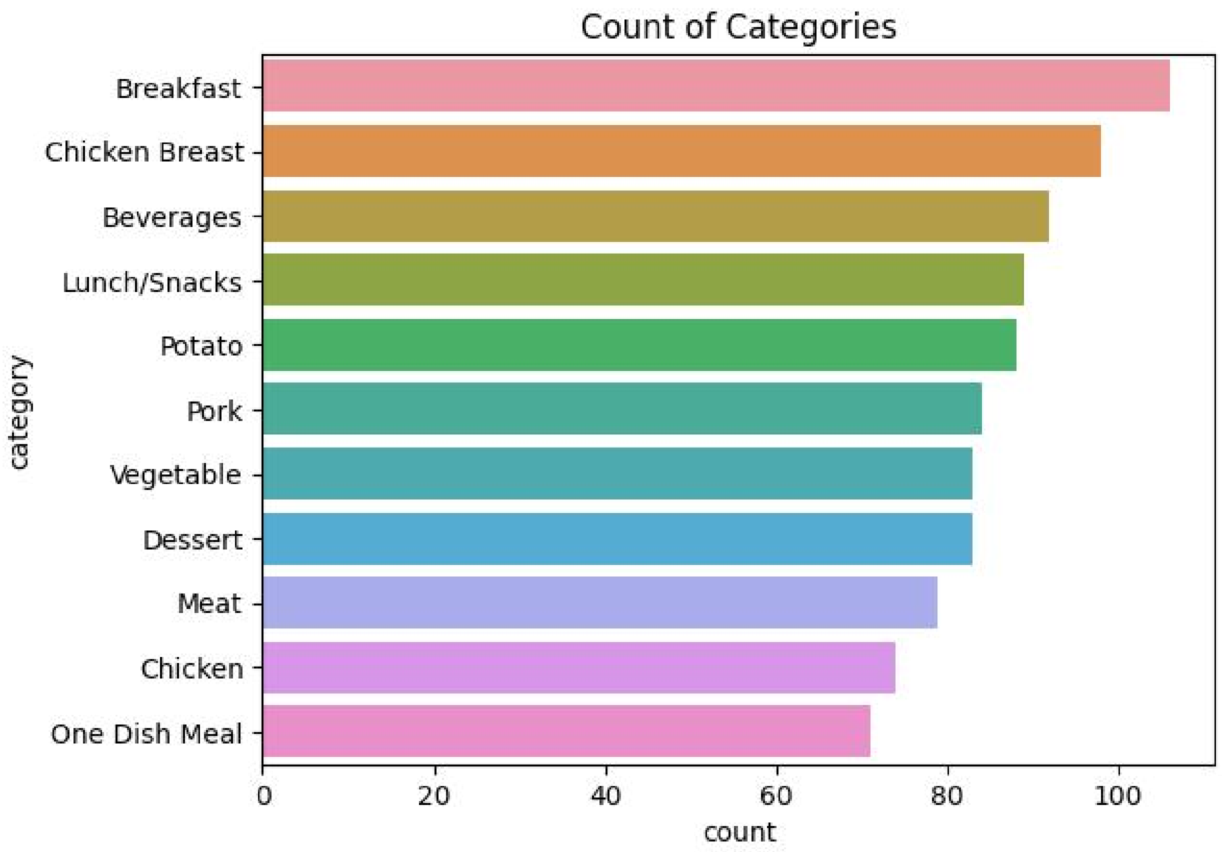
2- servings is not Numeric as there are “as a snack” extra part in only 3 rows:

- remove the substring “as a snack”

3- There are 52 missing values in columns ['calories', 'carbohydrate', 'sugar', 'protein'] which represents about 5% of the dataset:

- replacing them with the mean of data.

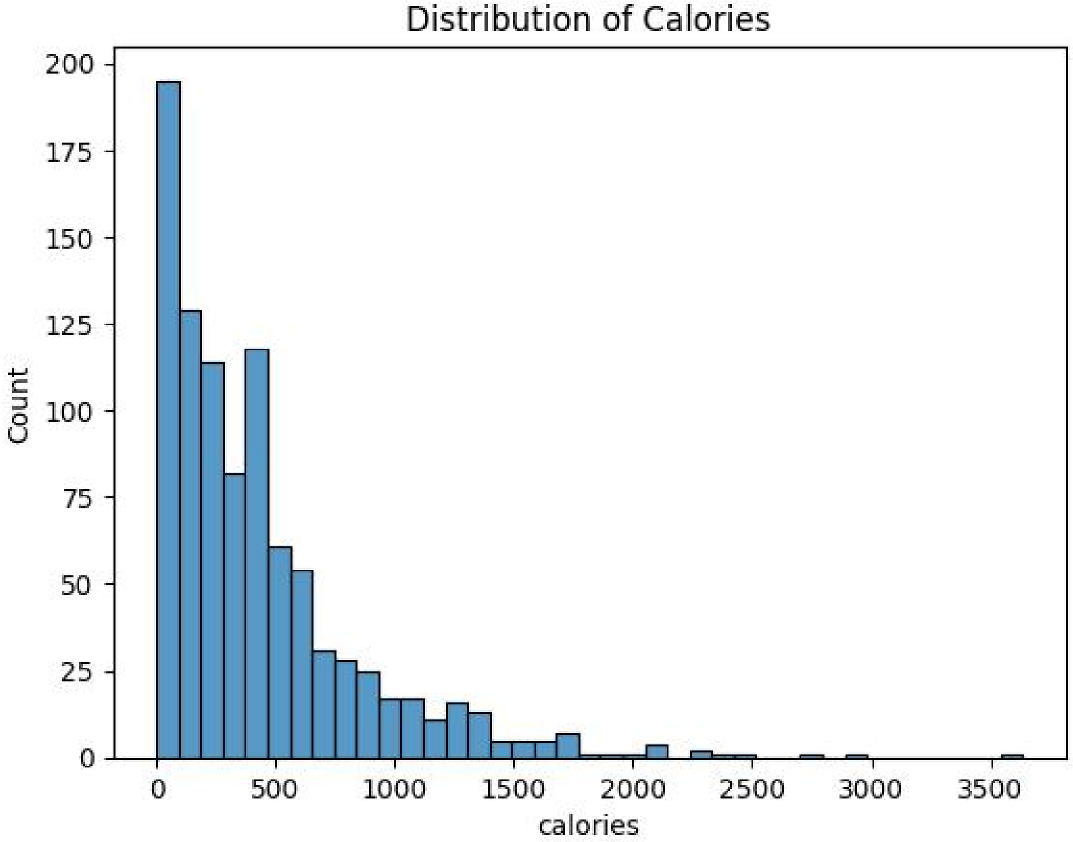
Data Visualization



- The number of Breakfast is the highest number posted.

- One dish meals are the least number posted

Data Visualization

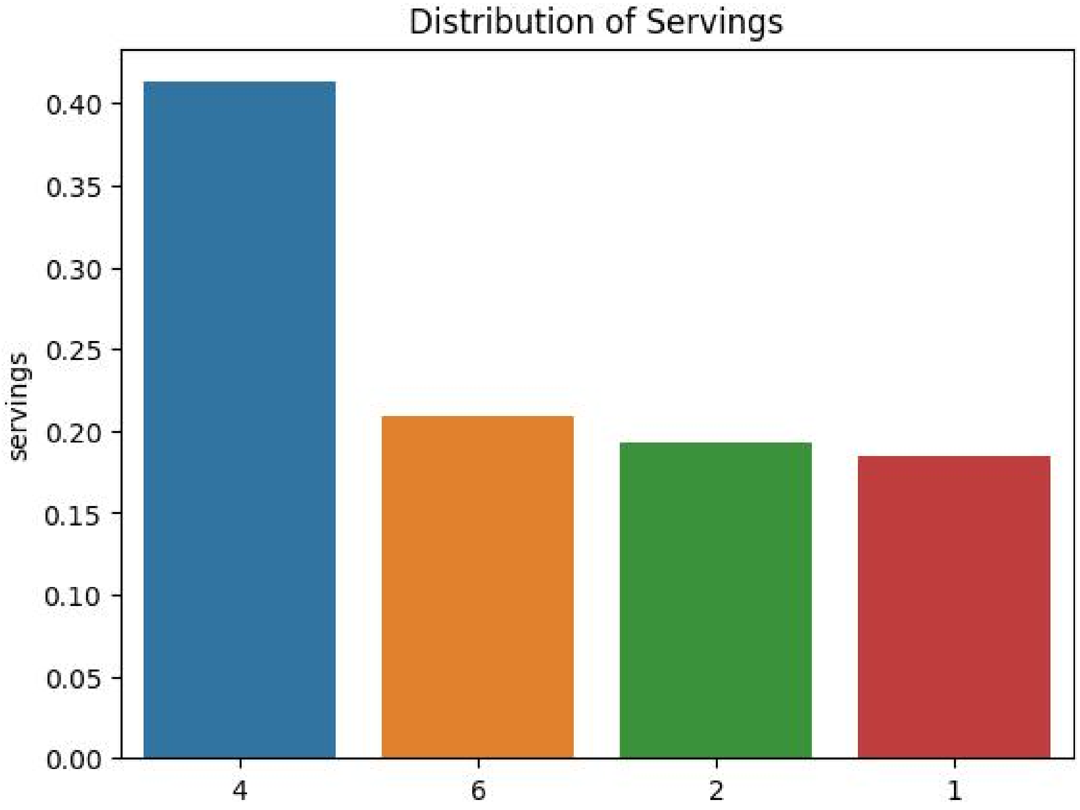


- The distribution of calories posted is right-skewed

- There are outliers

  [Data points > 2500cal]

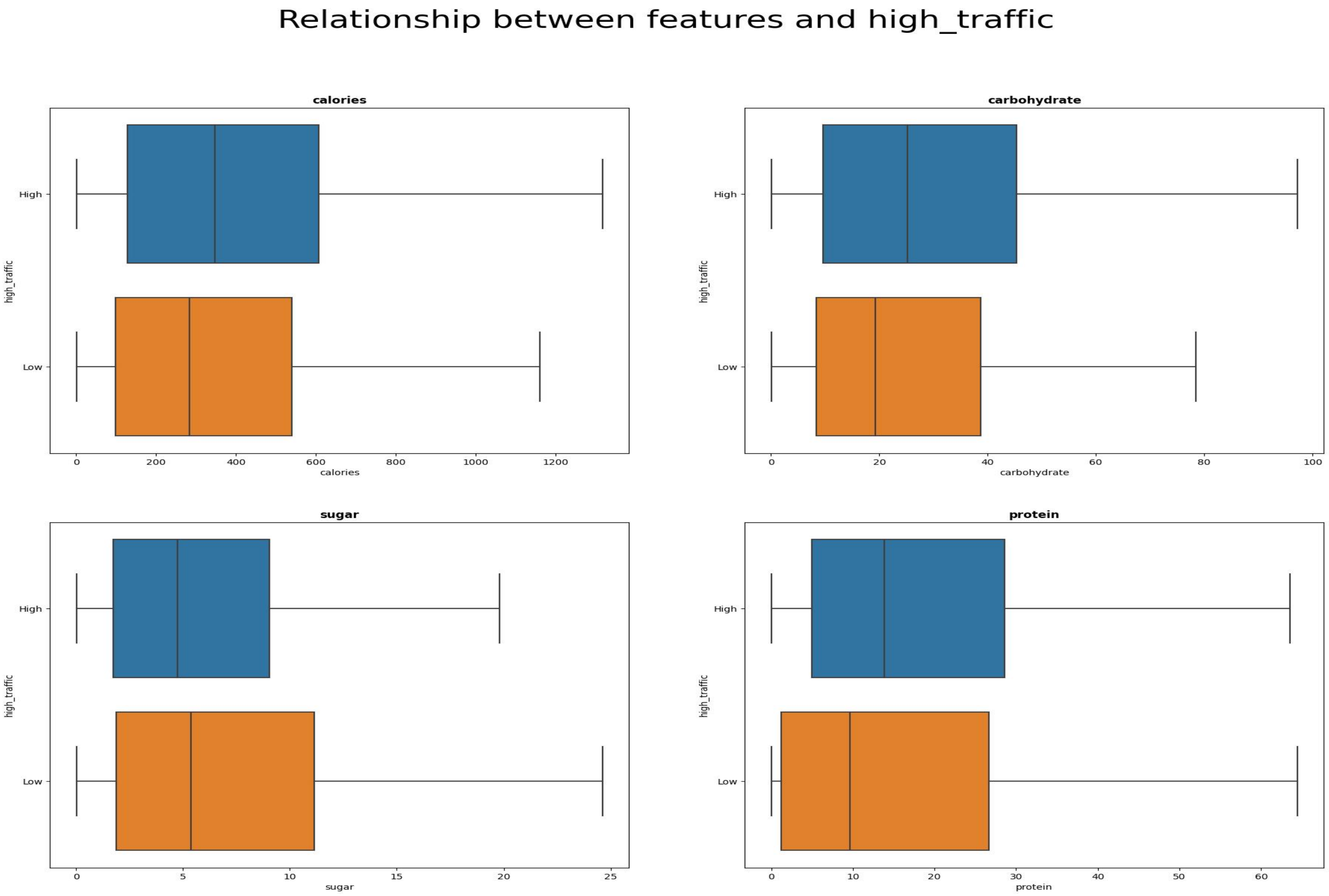
Data Visualization



The number of 4-servings is the highest number posted.

The number of 1-servings is the least number posted.

Data Visualization



We can summarize that

- The more calories the recipe, the more traffic
- The more carbohydrate the recipe, the more traffic
- The less sugar the recipe, the more traffic
- The more protein the recipe, the more traffic

Is this difference a real difference or by chance?

Subtract mean of high-traffic calories & mean of low-traffic calories = 65.37

The Hypothesis:

- Null Hypothesis (H0): mean for High-traffic calories <= mean for Low-traffic calories
- Alternative Hypothesis (H1): mean for High-traffic calories > mean for Low-traffic calories

Using T-test:

- p-value = 0.012 and threshold= 0.05
- so, we conclude (Mean of Calories for high traffic > Mean of Calories for Low traffic)

We can summarize that the more calories the recipe, the more traffic

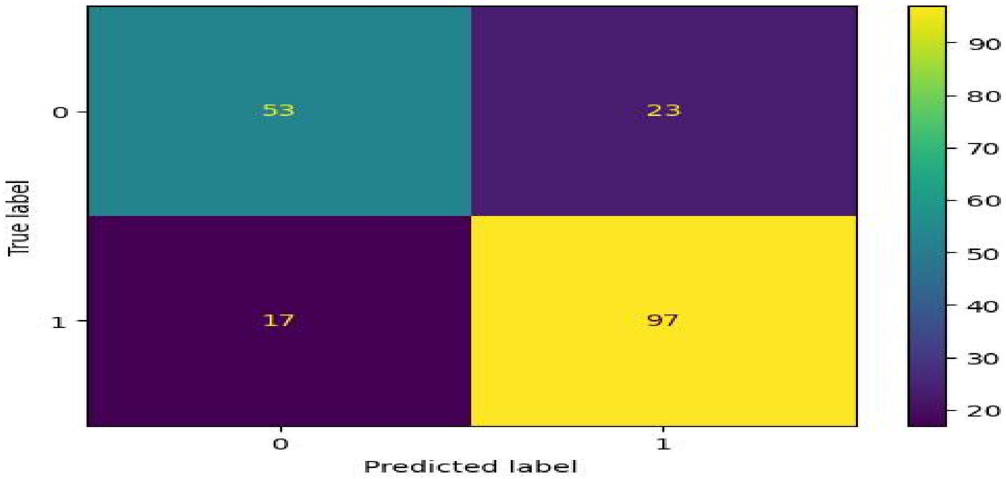
Model Development

This a classification problem (high / low)

- Fitting a comparison model:

Using RidgeClassifier, it can be significantly faster than LogisticRegression

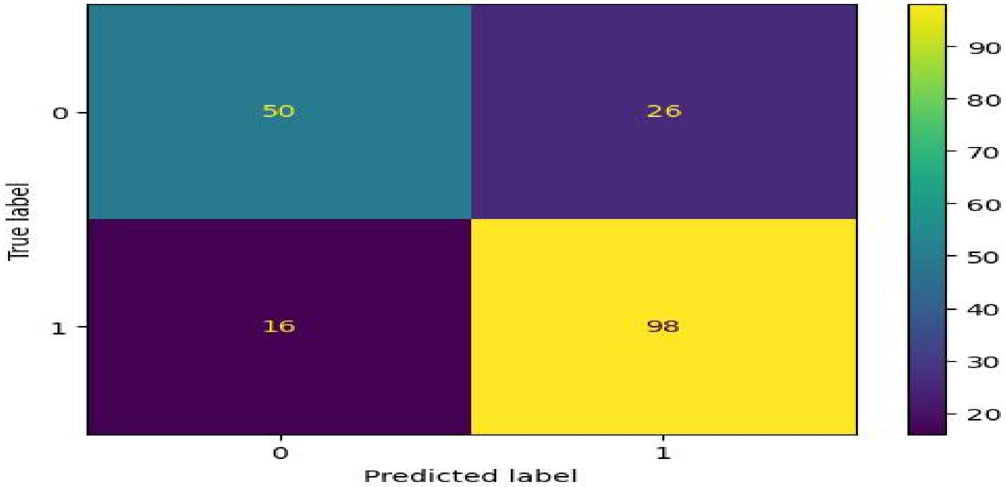
Score of model = 79%



- Fitting a comparison model:

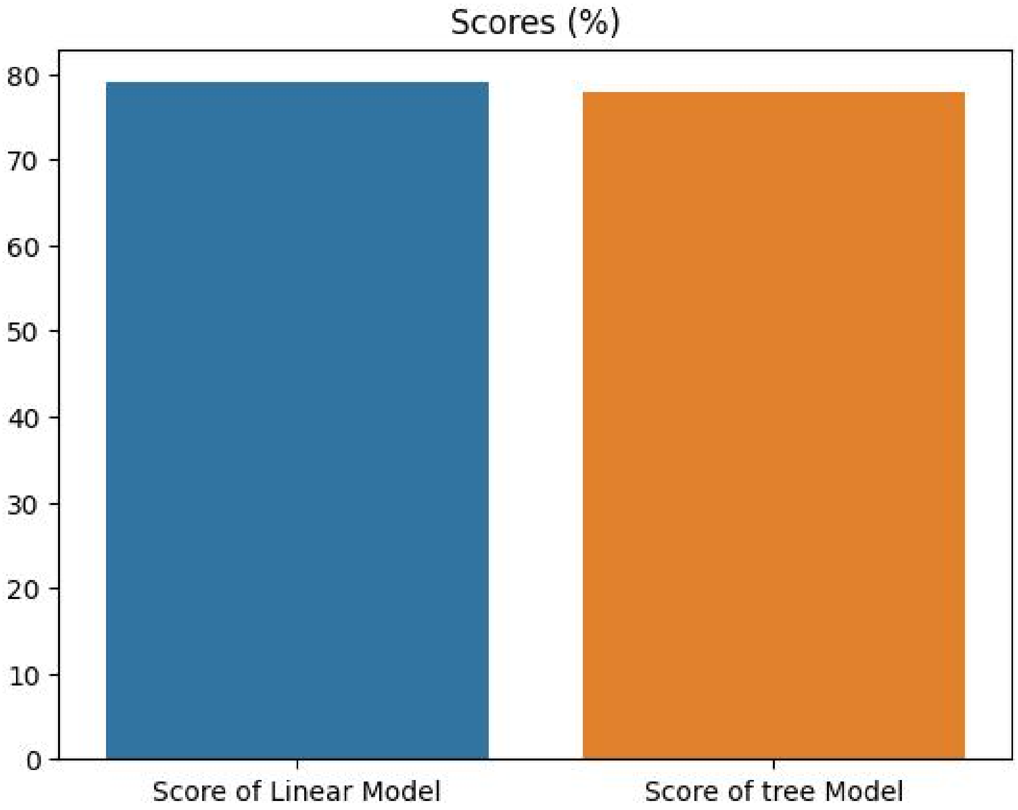
Using RandomForestClassifier, it can be more accurate than DecisionTree

Score of model = 78%

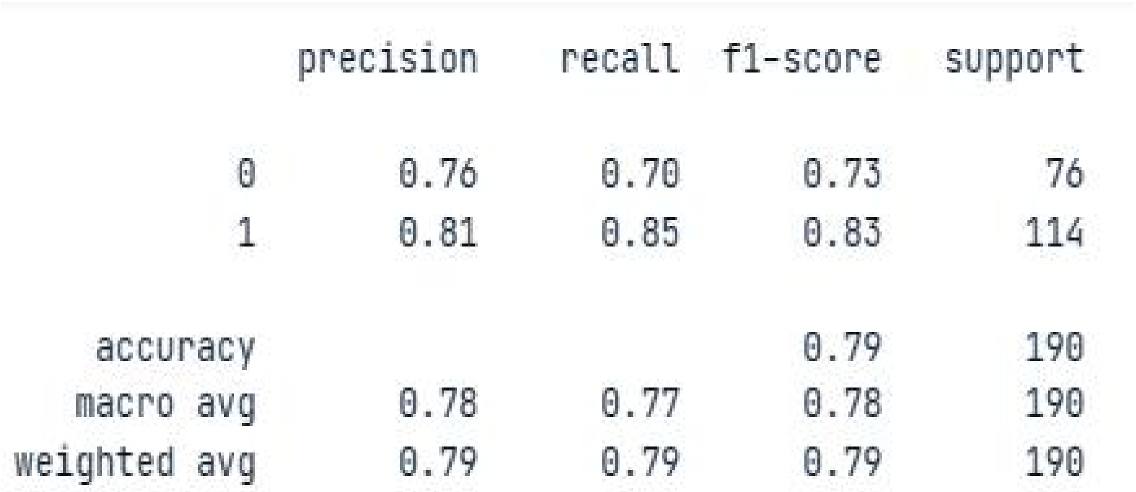


Model Evaluation

- The Linear Model more accurate than the Random Forest in both accuracy and recall score



Classification Report for Linear Model



```
              precision    recall  f1-score   support

           0       0.76      0.70      0.73        76
           1       0.81      0.85      0.83       114

    accuracy                           0.79       190
   macro avg       0.78      0.77      0.78       190
weighted avg       0.79      0.79      0.79       190
```

Business recommendation

- the business should implement the model since it can identify more than 80% of the high traffic generating recipes
- it is always a good practice to retrain the model with more data to improve its performance, in addition to capturing any changes in the customers behavior

I recommend:

- Iteratively improving the model over time with:
  - additional features
  - additional observations
  - screening more models and preprocessing techniques

- Utilize the current model as it’s shown to meet targets and provide business value, time to put the model into production.