Project Report

In the subsequent report, we hereby present a comprehensive analysis along with insights gleaned from our data science initiative, which was undertaken to tackle the quandary of identifying recipes that garner the highest traffic. Our primary objective encompassed the development of a predictive model capable of discerning whether a given recipe is likely to generate substantial traffic. Our project encompasses several pivotal phases, each of which will be elucidated in a sequential manner below. Each phase has been meticulously executed to ensure a profound grasp of the data, thereby facilitating the provision of actionable recommendations for the enterprise.

This endeavor is poised to prove instrumental for the Director of Data Science, enabling informed decisions rooted in empirical data and statistical analysis rather than relying solely on intuition.

Data validation:

To ensure the data's quality, a series of meticulous steps were implemented. Commencing with an examination of the data frame's variable data types, we proceeded to scrutinize the presence of NaN values.

A noteworthy observation arises from the non-random nature of the NaN values, which deviate from the Missing Completely at Random (MCAR) pattern. These absent values are primarily situated within the macronutrient variables, suggesting an omission in their calculation for certain recipes. Specifically, instances of NaN values for metrics such as calories, carbohydrates, sugar, or protein are consistently coexistent.

Encouragingly, these instances account for merely 5.5% of the dataset, signifying a relatively minor loss of information should we opt to omit them. Additionally, upon closer analysis, no discernible trend emerged, dispelling concerns over potential bias associated with missing data. Consequently, the distribution of these absent records when juxtaposed with the categorical variable remains largely stochastic, affirming that the data's missingness is, indeed, at random.

Notably, the remaining NaN values within the dataset pertain to the target variable "high\_traffic," which has been mapped to a Boolean representation, adopting the value "True" to denote recipes that yield substantial traffic.

Further meticulous exploration revealed an anomalous characteristic within the "servings" variable. Notably, this variable was identified as non-numeric, with three exceptions warranting rectification.

Exploratory analysis:

To ensure the data's high quality, the necessary step of outlier removal was undertaken. To identify these outliers, a boxplot was generated to visually represent the variable distributions.

.A graph of different colored squares

Description automatically generated with medium confidence

The presence of protein values surpassing 280 within a single meal deviates significantly from the norm, particularly when considering that a typical bodybuilder's daily consumption hovers around 200. This anomaly prompted further investigation, which revealed a similar aberration in the "carbohydrate" variable. These outliers were subsequently identified and subsequently removed. Further elucidation on this matter can be found in the "project.ipynb" file.

Similarly, the "calories" variable exhibited a parallel discrepancy, with certain recipes boasting an excessive 2000 calories or more. A thorough examination prompted us to establish a heightened threshold, whereby only those recipes exceeding 2500 calories were culled. It's important to note that this adjustment was informed by the presence of specific instances, such as one-dish meals, which could conceivably harbor such elevated caloric content.

A graph of a graph

Description automatically generated with medium confidence

To assess the equilibrium of the target variable and preempt any potential model bias, a histogram was employed. The analysis revealed a distribution whereby approximately 60% of records indicated high traffic, while the remaining 40% denoted low traffic.

* A graph with blue bars

  Description automatically generated

IMPORTANT:

During the exploration of traffic distribution in relation to recipe categories, a notable discovery emerged. It became evident that certain categories hold a greater appeal to clients compared to others.

A graph of different types of food

Description automatically generated

Our business, driven by the mission to "ensure you and your family are attaining good health," is evidently operating in alignment with its intended purpose. An intriguing trend emerges as clients exhibit heightened enthusiasm for health-conscious meal options, particularly within categories such as vegetables and potatoes, while displaying a decreased affinity for Beverages.

With this insight in mind, a strategic shift may prove beneficial by discontinuing the production of Beverages and reallocating resources towards the expansion of recipes within the vegetable and potato categories.

Similarly, an assessment was conducted to ascertain whether the volume of servings influences traffic. Remarkably, the analysis unveiled a distinct lack of discernible impact from serving size on client engagement. This intriguing finding suggests that allocating resources based on serving quantity holds minimal significance. Instead, our focus should shift to the provision of recipes independent of their serving scale.

Top of Form

A graph of a bar chart

Description automatically generated with medium confidence

Model Development:

The focal point of this challenge can be characterized as a binary classification endeavor, entailing the task of categorizing data into two distinct outcomes: "high\_traffic" and "no\_high\_traffic."

To establish a solid starting point, a logistic regression model was adopted and implemented using the sklearn library.

However, given that the performance of the logistic regression model fell short of expectations, a comparative evaluation was deemed necessary. To this end, alternative candidate models were explored, including the linear support vector classifier and the random forest classifier.

Despite this exploration, neither the linear support vector classifier nor the random forest classifier exhibited superior performance in comparison to the logistic regression model. Consequently, an additional candidate ensemble machine learning algorithm, namely the voting classifier, was introduced, amalgamating the predictive power of the individual models. The ensuing synthesis of their performance showcased notable enhancement, surpassing that of the previously considered models. A detailed presentation of these outcomes is provided below.

A black screen with white numbers

Description automatically generated

In addressing this challenge, we adopt the premise that both Precision and Recall bear equal significance. Hence, our attention centers on the F1 score, which adeptly synthesizes these two metrics into a single evaluation criterion.

The pronounced accuracy observed in the Voting Classifier's outcomes underscores the potency of amalgamating multiple models' predictions, yielding an elevated overall performance. Notably, this classifier strikes an appealing balance between precision and recall, rendering it a compelling candidate for the classification task at hand.

While the logistic regression model also exhibited commendable performance, its sensitivity to variations in training data sampling poses a potential limitation. This stands in contrast to the Voting Classifier, which demonstrates robustness against such fluctuations.

In light of these considerations, the Voting Classifier stands as the chosen model, fortified by its capacity to harmoniously blend precision, recall, and accuracy in a manner that aligns closely with the task's requirements.

Metric for the business to monitor:

As demonstrated through our earlier graphical analyses, it is evident that the traffic patterns are notably influenced by the distribution of recipe categories. Consequently, a prudent approach to monitoring the business would involve a dedicated focus on:

**The Traffic-Category Distribution:**

The analysis of recipe distribution across various categories underscores their significant impact on traffic generation. At present, there exists a notable emphasis on Breakfast and chicken breast categories, despite their comparatively lower customer attraction when juxtaposed with the vegetable, Potato, Pork, and One Dish Meal categories – which consistently yield higher levels of traffic. **Traffic per 100 recipes:**

This key performance indicator (KPI) reflects the proportion of recipes that yield high traffic per 100 recipes. As it stands, this metric currently stands at approximately 59.7%. Recognizing the significance of elevating this metric, implementing strategic enhancements becomes imperative for the progression of the business.

Summary:

Several areas within the business necessitate improvement, and we propose the following strategic actions:

**Dropping the Beverage Recipe Category:**

Considering the observed trend, discontinuing the beverage recipe category appears to be a prudent step. Redirecting resources away from this category could lead to more effective utilization of available assets.

**Reducing Emphasis on Servings per Recipe:**

Our analysis indicates that the number of servings per recipe has minimal impact on traffic generation. Thus, shifting the focus away from this aspect could lead to improved resource allocation.

**Enhanced Focus on Select Categories:**

Given the consistently higher traffic generation, channeling greater attention towards vegetable, Pork, Potato, and One Dish Meal recipes holds promise for boosting overall business performance.

**Optimizing Macronutrient Calculation:**

Our models suggest that the calculation of macronutrient content may not significantly impact traffic. As such, reconsidering the allocation of budget and resources for this aspect could lead to more efficient resource management.

By strategically addressing these suggested areas of improvement, the business stands to achieve heightened performance and better alignment with customer preferences and market trends.