TopUp Mama

Pamela Mwirigi

Abstract

This report explores TopUp Mama's data. TopUp Mama is a restaurant partner in Africa.

The report explores how to handle the above tasks through harnessing the company's data and employing machine learning models to draw insight for the company.

It applies the principles of predictive algorithms, recommender systems, classifiers, and clustering to develop models for the business. It employs different algorithms, evaluation metrics for comparing models, and imputation methods for handling data.

Introduction

Customer retention

The cost of attracting new customers is much higher than the cost of retaining customers. Therefore, businesses are tasked to pay attention to their customers. A majority of customers will not express their sentiment prior to leaving a business and therefore businesses have turned to ML to predict customer churn based on data.

Classify customers'

Classification of customers is popularly known as Customer Segmentation which is important for businesses to understand their audience. This way a company can accurately target its audience through well-curated campaigns based on their profiles, interests, demographics, etc. A popular way to accomplish this is through unsupervised learning.

Product recommendations

Recommender systems are employed to rate the preference a user has for a particular item. They're employed by almost every tech company now. For instance, on YouTube, they give you video recommendations, on Reddit they give thread recommendations, on Netflix, movie recommendations, on LinkedIn, connections, etc. They are classified into three types which include simple, content-based, and collaborative filtering engines.

Revenue Optimization

Revenue optimization is coupled with pricing optimization which is the lever for revenue growth within companies. Machine learning is well suited for price optimization as it is able to handle complex features and generalize them to new situations.

ORIGINAL DATASET

Our original dataset is 91 dimensional with 66673 observational features.

The features include:

'Customer_ID', 'Last_Used_Platform', 'Is_Blocked', 'Created_At', 'Language', 'Outstanding_Amount', 'Loyalty_Points', 'Number_of_Employees', 'Upload_restuarant_location', 'Task_ID', 'Order_ID', 'Relationship', 'Team_Name', 'Task_Type', 'Notes', 'Agent_ID', 'Agent_Name', 'Distance(m)', 'Total_Time_Taken(min)', 'Pick_up_From', 'Start_Before', 'Complete_Before', 'Completion_Time', 'Task_Status', 'Ref_Images', 'Rating', 'Review', 'Latitude', 'Longitude', 'Tags', 'Promo_Applied', 'Custom_Template_ID', 'Task_Details_QTY', 'Task_Details_AMOUNT', 'Special_Instructions', 'Tip', 'Delivery_Charges', 'Discount', 'Sub_Total', 'Payment_Type', 'Task_Category', 'Earning', 'Pricing', 'Unnamed: 34', 'Unnamed: 35', 'Order_ID', 'Order_Status', 'Category_Name', 'SKU', 'Customization_Group', 'Customization_Option', 'Quantity', 'Unit_Price', 'Cost_Price', 'Total_Cost_Price', 'Total_Price', 'Order_Total', 'Sub_Total', 'Tax', 'Delivery_Charge', 'Tip', 'Discount', 'Remaining_Balance', 'Payment_Method', 'Additional_Charge', 'Taxable_Amount', 'Transaction_ID', 'Currency_Symbol', 'Transaction_Status', 'Promo_Code', 'Customer_ID', 'Merchant_ID', 'Store_Name', 'Pickup_Address', 'Description', 'Distance(m)', 'Order_Time', 'Pickup_Time', 'Delivery_Time', 'Ratings', 'Reviews', 'Merchant_Earning', 'Commission_Amount', 'Commission_Payout_Status', 'Order_Preparation_Time', 'Debt_Amount', 'Redeemed_Loyalty_Points', 'Consumed_Loyalty_Points', 'Cancellation_Reason', 'Flat_Discount', 'Checkout_Template_Name', 'Checkout_Template_Value'

Data Preparation:

- Feature selection,
- training set and test set selection

General Observations:

The data is extremely sparse with columns spanning 66673 features having more than 80% of their data missing. Some of these columns happened to be critical to computations yet they did not meet the threshold for inclusion.

Conclusively, as this data is not rare making it critical to the organization, it indicates a lack of proper data collection techniques or in the very least methods that are not effective.

In order to serve the purpose of this exploration, I have adapted a leaner dataset for every task. Narrowing down to 33 for customer retention and revenue optimization, two for a recommendation system, and 8 for customer segmentation.

Missing Data

Because a significant portion of data is missing, simple imputation is used to fill in the data. This is accomplished through imputing mean, mode, and random samples from the training set.

This substitution, however, raises the risk for overfitting which is observed in our modeling process.

Feature selection

I have selected features based on their relevance to the task and excluded highly correlated variables.

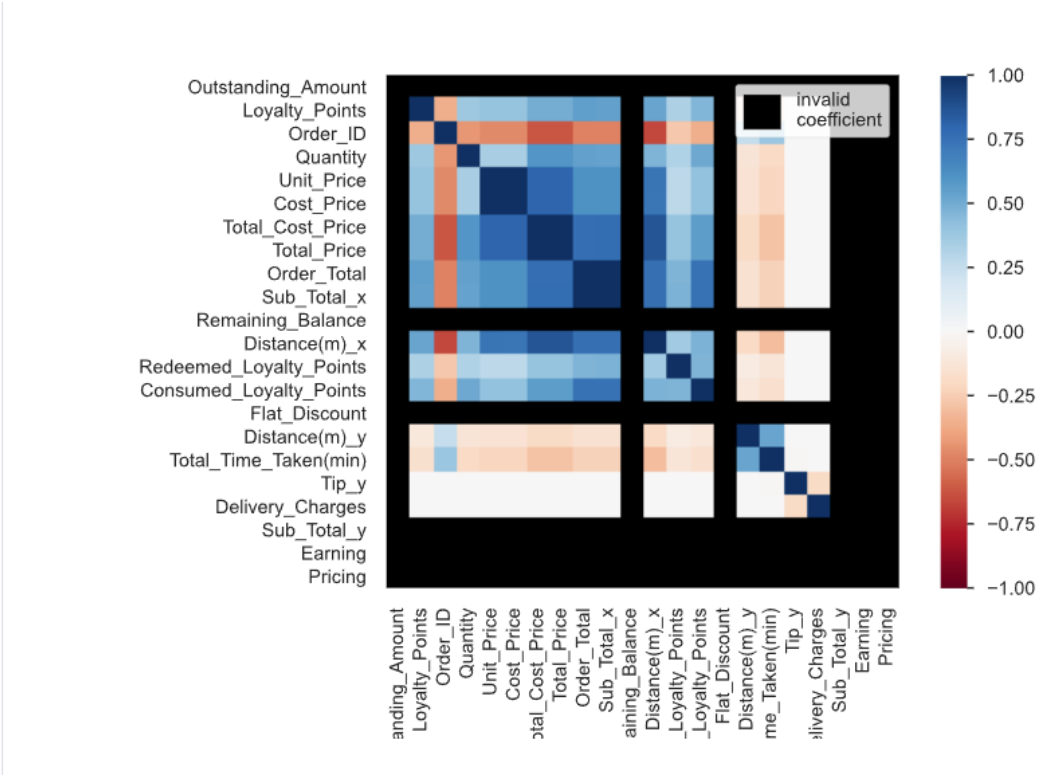
Training set and Test set Selection

Where necessary I have employed a training and test split of 0.2 from the original dataset. I have also replaced categorical variables with numerals by employing a label encoder over manual encoding.

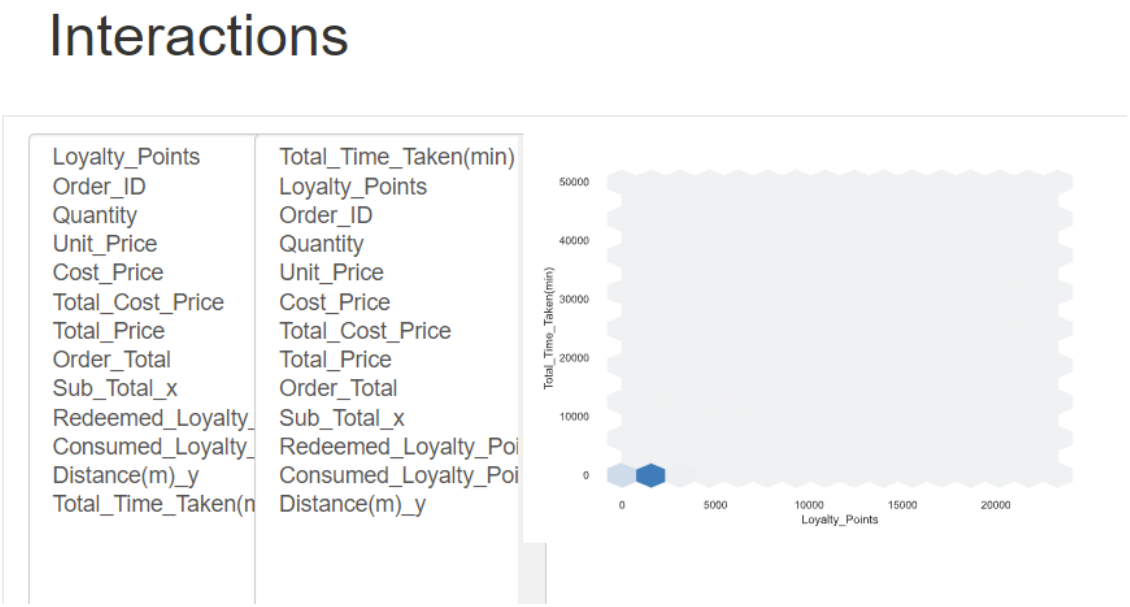
First rows

	Last_Used_Platform	Is_Blocked	Language	Outstanding_Amount	Loyalty_Points	Order_ID
0	WEB	0.0	en	0.0	68.0	11155410
1	WEB	0.0	en	0.0	68.0	11155410
2	WEB	0.0	en	0.0	68.0	11155410
3	WEB	0.0	en	0.0	68.0	11138864
4	WEB	0.0	en	0.0	55.0	11253217
5	WEB	0.0	en	0.0	55.0	11253217
6	WEB	0.0	en	0.0	55.0	11253217
7	WEB	0.0	en	0.0	55.0	11216794
8	WEB	0.0	en	0.0	55.0	11216794
9	WEB	0.0	en	0.0	55.0	11216794

Correlations:



Interactions:



Methods

I have employed logistic regression, linear SVM, random forest, and gradient boosting for customer retention.

In customer segmentation, I have used K-means to cluster the data and further advanced it using a silhouette coefficient which evaluates the quality of the clusters created by an algorithm. The higher the score, the better the model.

In product recommendation, I have used a content-based recommender as it suggests items based on history. The rankings were too few to use to build a collaborative filtering system. In the content-based recommender, I have used cosine similarity to measure the distance.

In revenue optimization, I have used different predictive algorithms to determine price.

Algorithm Selection

The algorithms chosen are the best suited for the tasks at hand. However, we find that the dataset gets messy and the algorithms in some cases tend to overfit as the analysis is only as good as the data. We are able to distinguish between the false positives and false negatives and to effectively lower them.

There is the use of different features to better tune our models. For instance, PCA, Principal Component Analysis is used to reduce the model's complexity and to reduce random noise. Its purpose is to reduce dimensionality using Singular Value Decomposition.

Conclusion and Discussion

Examining the present customers and learning the customers allows the business to identify segment possibilities and to explain them but after it is critical to research the segments and understand how to transform products and approaches for the said segments. Having proper pricing procedures is also key to ensuring that businesses are capable of maximizing their revenues.