Coupon Purchase Prediction

Predict which coupon a customer will buy

Abraham G.K(20BDA20), Rakshith Kumar K N(20BDA47), Chandan J.R(20BDA51)

Background:

In E-commerce contexts, issuing discount shopping coupons is a common approach to encourage sales. However, it is unknown whether a client will utilise the voucher. It is perspective to increase the coupon usage ratio in order to boost the coupon usage ratio. It is important to estimate the likelihood of a user redeeming a discount weather the person is goanna use coupon / weather the customer is goanna buy coupon.

We'll investigate coupon utilisation likelihood prediction in this section. We treat the problem as a binary classification problem and utilise machine learning to solve it. Methods for analysing consumers' coupon usage behaviour are being learned. We carry out a thorough research of coupon usage patterns among consumers

Coupon Purchase Prediction's data is available on the Kaggle platform.

link: https://www.kaggle.com/c/coupon-purchase-prediction

Introduction:

E-commerce platform vendors may provide major promotions (e.g., discounts or cash coupons) on specified days or during specific vacations in order to attract a high number of new consumers.

Discount marketing and couponing are common promotional strategies for recruiting new consumers as well as keeping and strengthening existing customer loyalty. The ability to measure a consumer's propensity to use coupons and estimate redemption behaviour are crucial factors in determining the success of a marketing campaign.

Promotions for the client are sent out via a range of channels, including email, notifications, and social media. In a number of these promotions, coupons are issued for a single product or a range of items. The retailer's marketing team would like to be able to predict whether

customers would redeem coupons received across channels, allowing them to design more precise and focused coupon structures and generate more accurate and targeted marketing techniques.

Literature Survey:

To enable real-time web marketing, there are two types of algorithms for predicting the probability of unknown customers purchasing." The suggested method is divided into two parts: gathering purchase trends and assessing the likelihood of purchasing. Purchase probability provides real-time web marketing opportunity by calculating an unknown consumer's potential of purchasing, whereas purchasing trend provides marketing clues to web marketers. Because unknown consumers are participating in the target marketing and the buying navigation pattern is developed, these approaches may be significant for real-time online marketing that likes shortcuts, Brand advise, and best customer reward.

In this study, we analyse the topic of coupon purchase prediction as a binary division problem, and we utilise machine learning algorithms to anticipate client coupon usage behaviour." We do a large-scale data analysis using Ponape's data provided via Kaggle.

Some users thought that the Random Forest and XgBoost were better for building models, and that the Random Forest with data and cutoff probabilities was one of the models they preferred. However, the modelling has several implementation limitations and potential model hazards. They couldn't guarantee the model's stability.

Users are rewarded with a coupon code if they achieve the high-time fair of the lowest operational cost." The dependency of the locational marginal cost and the market is formed by an artificial neural network. The model's results reveal that customer coupon demand response leads to cutting the large, resulting in significant economic savings and loss reduction, as well as a reduced price for the product being purchased due to the coupon code.

Hypothesis/Aim of the work Do not make claims without supporting data, reference or hypothesis testing

CHI-SQUARE TEST

To test if two variables in a contingency table are connected, the Chi-square Test of Independence is used. In a broader sense, it examines whether categorical variable distributions differ from one another.

To analyse the frequency table (i.e., contingency table) created by two categorical variables, the Chi-square test of independence is utilised.

The term - represents the chi-square statistic.

The Degrees of Freedom are represented by C, the Observed Frequencies are represented by Oi s, and the Expected Frequencies are represented by Ei s. [Degree of freedom = (number of rows - 1) * (number of columns -1)].

There is no link between the categorical variables, according to the null hypothesis.

Alternative hypothesis: The categorical variables have a link.

The Level of Significance () is crucial in this case.

If the p-value is greater than: Significant result, reject the null hypothesis (H0)

If the p-value is greater than, the null hypothesis is not rejected (H0)

Aim of the work

Primary goal: Predict which coupons a consumer will buy in a particular period of time based on previous purchase and browsing activity. From July 2011 to June 2012, the data comprises a year's worth of transactional data for 22873 users. Over the next week, predictions will be made on 310 additional coupons, each with its own set of qualities. The goal is to guess which of these 310 coupons will be purchased by each of the 22873 people.

Conduct exploratory analysis on transactional data for 22,873 Ponpare users as a secondary aim.

Motivation: The generated models will be utilised to increase the ability to predict which coupons a consumer will purchase, ensuring that they do not miss out on their next favourite item.

Methods and Materials

Dataset

Ponpare, a coupon site that we found through Kaggle, provided us with user transactional data. From July 1, 2011, through June 23, 2012, the training set comprises 359 days of consumer activity. The test set runs from June 24 to June 30, 2012, one week following the training set, and comprises seven days of client engagement.

Variables in datasets

User list.csv \User list.

The master list of users in the dataset is called csv. In user list.csv, there are a total of 22,873 records.

Coupon list train.csv and Coupon list test.csv.

Coupon list train.csv is a master list of coupons that make up the training set. Coupon list train.csv has a total of 19,413 entries.

The master list of coupons that are considered part of the test set is coupon list test.csv. Coupon list test.csv has a total of 310 records. The 310 coupons will be used to make predictions for this project.

Coupon visit train.csv.

The viewing log of users browsing coupons throughout the training period is contained in the file coupon visit train.csv. There are 2,833,180 records in coupon visit train.csv.

Coupon detail train.csv.

The coupon detail train.csv file contains a purchase record of users who purchased coupons throughout the training session. Coupon detail train has a total of 168,996 records.

Coupon area train.csv and Coupon area test.csv are two different types of spreadsheets.

The file coupon area train.csv contains a list of coupons and area locations that make up the training set. Coupon list train.csv has a total of 13,8,185 entries in it.

The file coupon area test.csv contains a list of coupons and their locations in the test set. The coupon area test.csv file has a total of 2142 entries. The 2142 coupons will be used to generate predictions for this project.

Packages: Sklearn (scikit-learn), NumPy, Pandas, matplotlib, seaborn, xgboost, Kera's, Seaborn Algorithm's used: Random Forest, XGBoost, ANN for recommendation system

Study Design:

Data Processing:

We used five different datasets in this research and combined them into a single data frame. Then we made a news column called 'Article' for the text, which will be the combination header and content. 1 was replaced to true and 0 was replaced to fake in the Label field.

Step1 Download datasets:

Step2 Data transformation (deriving and casting)

Translate Japanese text in **PREF_NAME** into English and save it into new variable **PREF_NAME_en**:

Data casting - changing type of the variables for further analysis

Step3 Data deriving

Translate Japanese text in **CAPSULE_TEXT**, **GENRE_NAME**, **large_area_name**, **ken_name**, **small_area_name** into English and save it into new corresponding variables **CAPSULE_TEXT_en**, **GENRE_NAME_en**, **large_area_name_en**, **ken_name_en**, **small_area_name_en** in training and test data sets:

Remove all unwanted columns.

Remove all records with missing values.

The weight parameters of the variables were condensed to the interval between one and zero, and all NAs were substituted with one.

The cosine similarity method was utilised. By comparing the similarity of each coupon in the test dataset to the coupons purchased by the user previously in the training dataset using matrix analysis, cosine similarity is calculated. All 310 coupons in the test dataset are rated for each user based on their similarity to the same user's previous purchases in the training dataset. The top ten highest-scoring coupons for each user are picked to appear in the user's suggestion list. The most significant part of the cosine similarity model was creating a weight matrix by giving various weights to different features. I selected the most significant properties, which are coupons category, discount price, pricing rate, useable date, and location of the business or service where the user may redeem the coupon, using a trial-and-error technique.

Model ‘s Used:

Random Forest

Bagging is another use of ensemble trees, in which several large trees are fitted to bootstrap re-sampled copies of data then categorised by majority vote. RF improves on Bagging by de-correlating the trees. Following each tree split, a random sample of features is chosen, and only these are used in the next split. Once again, the results are based on the majority vote of the single trees. By applying a large number of classifiers, bagging and RFs improve on the shortcomings of non-ensemble DTs, such as robustness and over-fitting. They train faster than boosted trees, but it takes longer to anticipate. On the other hand, RFs and bagging still rely on feature engineering and are unable to explain time dependencies.

XGBoost

XGBoost is a distributed gradient boosting toolkit that has been tuned for efficiency, flexibility, and portability. It uses the Gradient Boosting framework to create machine learning algorithms. XGBoost is a parallel tree boosting (also known as GBDT, GBM) algorithm that solves a variety of data science issues quickly and accurately.

Artificial Neural Network

One of the most fundamental Neural Network models is the Artificial Neural Network. Only forward data passes through the network, from the input nodes to the hidden nodes, and finally to the output nodes. Layers of nodes are created, with each layer's node linking to the nodes of the preceding layer. The connections that are created during the learning phase, such as via back-propagation, are given different weights.

Every node has an activation function that determines when the node should fire. The output is the probability of each class, which adds up to one. If there are enough hidden units in an ANN, it can approximate any function. From a statistical viewpoint, ANNs execute nonlinear regression. Such neural networks have two drawbacks: long training durations and limited comprehensibility. They demand feature engineering as well as hyper-parameter tuning. They have, nonetheless, proved

Evaluation of model: In this we anticipate the output for our test data and use y test to assess the expected results.

Model Evaluation Matrix:

The Mean Squared Error (MSE)

Mean Squared Error is a measure of how accurate (MSE). MSE is a widely used and straightforward statistic that accounts for a small change in mean absolute error. Finding the squared difference between the actual and anticipated value is defined as mean squared error.

So, we found the absolute difference above, and we found the squared difference below.

What does the MSE truly stand for? It denotes the difference in squared values between actual and expected values. The benefit of MSE is that we conduct squared to prevent the cancellation of negative terms. MSE (Mean squared error) is a regression assessment statistic.

The Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is a measure of how accurate a (RMSE).In regression issues, the most commonly used assessment measure is the root mean square error (RMSE). It is based on the premise that errors are random and have a normal distribution.

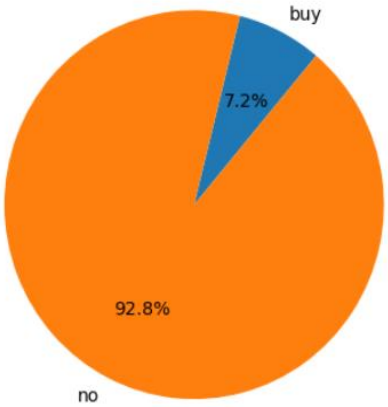
The strength of the metric ‘square root' allows it to display significant number variances. The metric’s ‘squared' character aids in the delivery of more reliable findings by preventing the cancellation of positive and negative error values. To put it another way, this metric accurately depicts the likely amount of the error term.

It avoids the use of absolute error numbers in mathematical computations, which is extremely undesirable. Reconstructing the error distribution using RMSE is believed to be more trustworthy when we have more samples.

Outlier values have a significant impact on RMSE. As a result, before utilising this measure, make sure you've eliminated any outliers from your data collection.

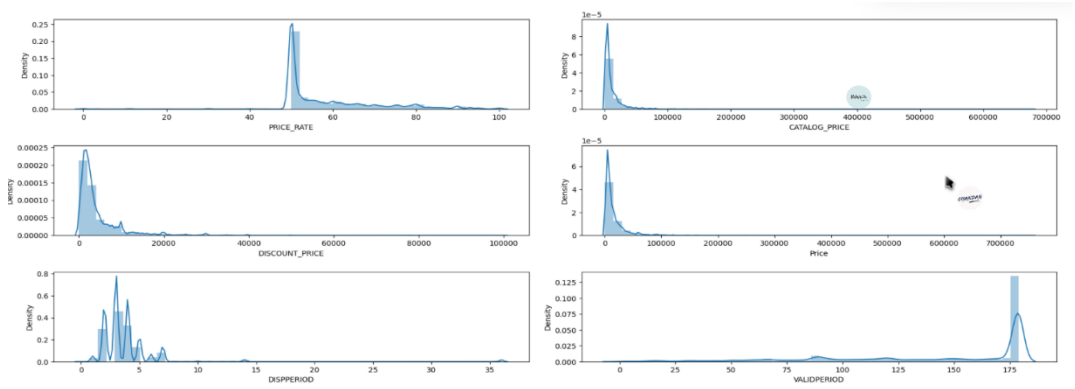
When compared to mean absolute error, RMSE gives more weight to small mistakes and penalises big ones. is a measure of how accurate a (RMSE). In regression issues, the most used assessment measure is the root mean square error (RMSE). It is based on the premise that errors are random and have a normal distribution.

Exploratory Data Analysis:

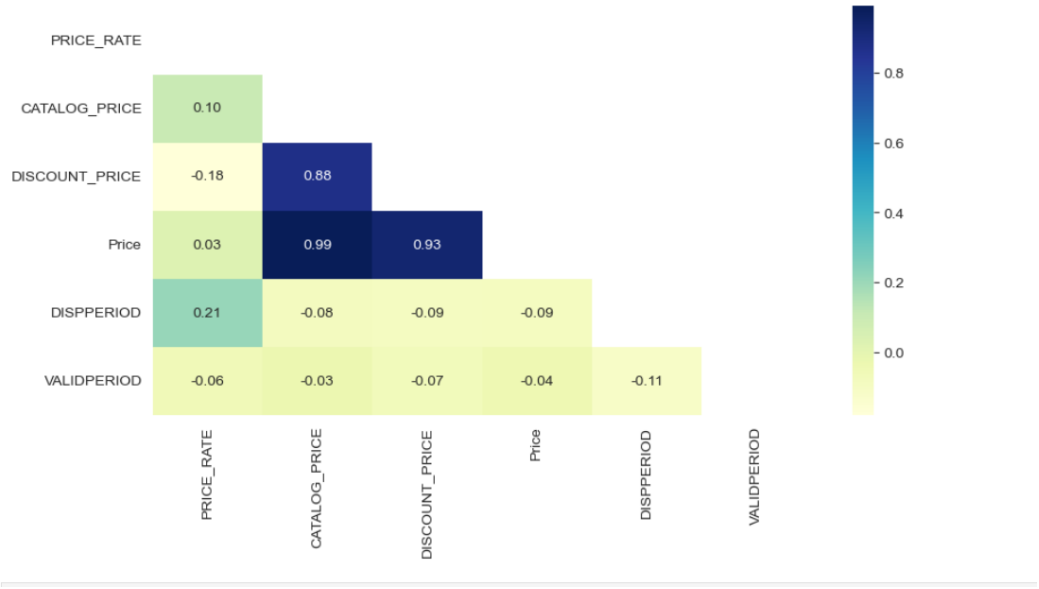


From the above plot We'll see how many buyers will buy things utilising coupons by considering features like "Price Rate”, “CatLog Price", "Discount Price", "Actual Price" ,"Disp Period" and "Valid Period."

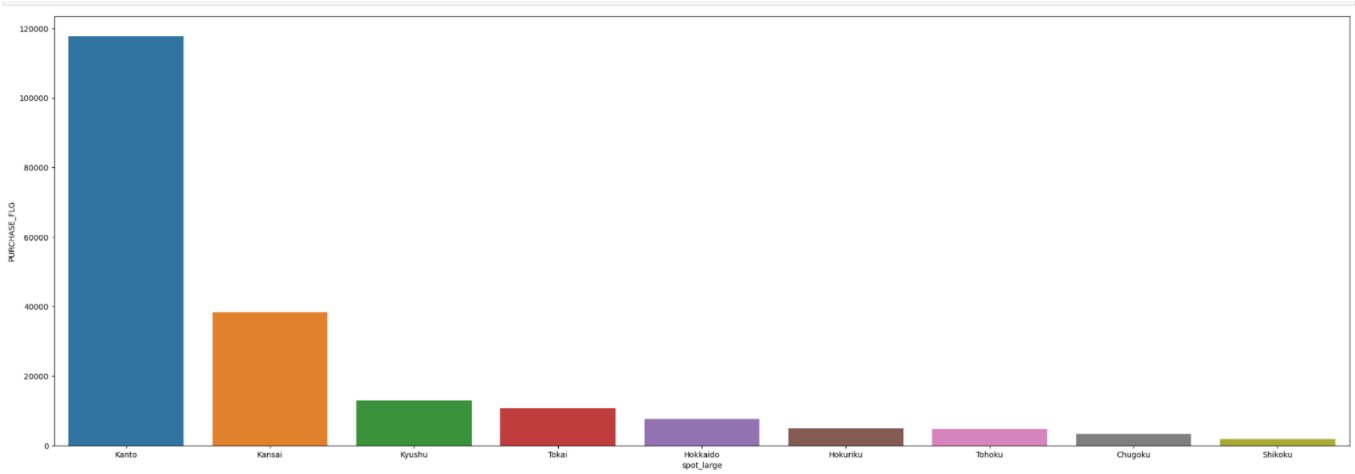
That is 7.2% of customers are buying products using coupons where other 92.8% customers are buying products without using coupons.



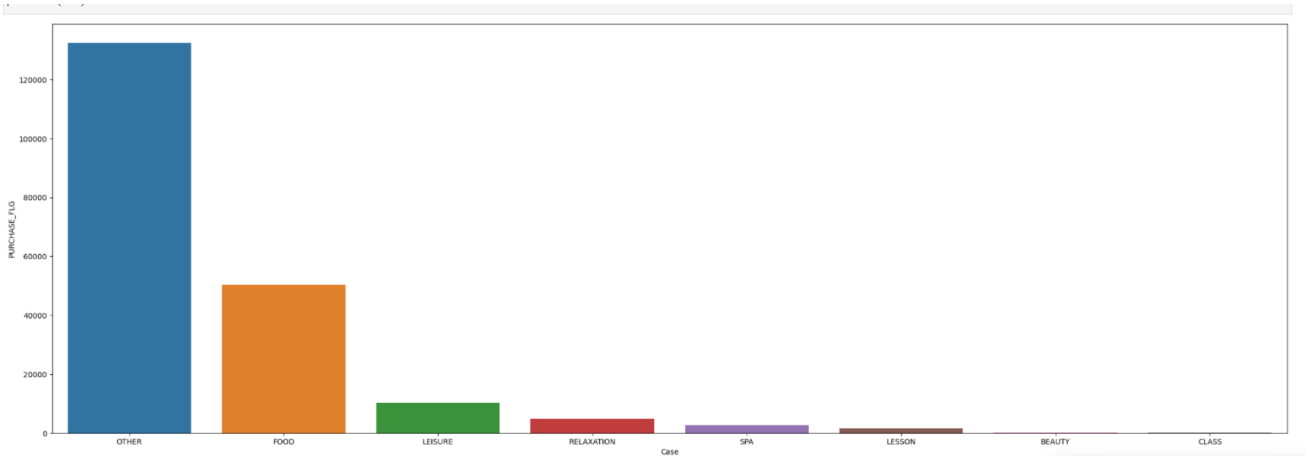
The distributions of Price are depicted in the graphs above.



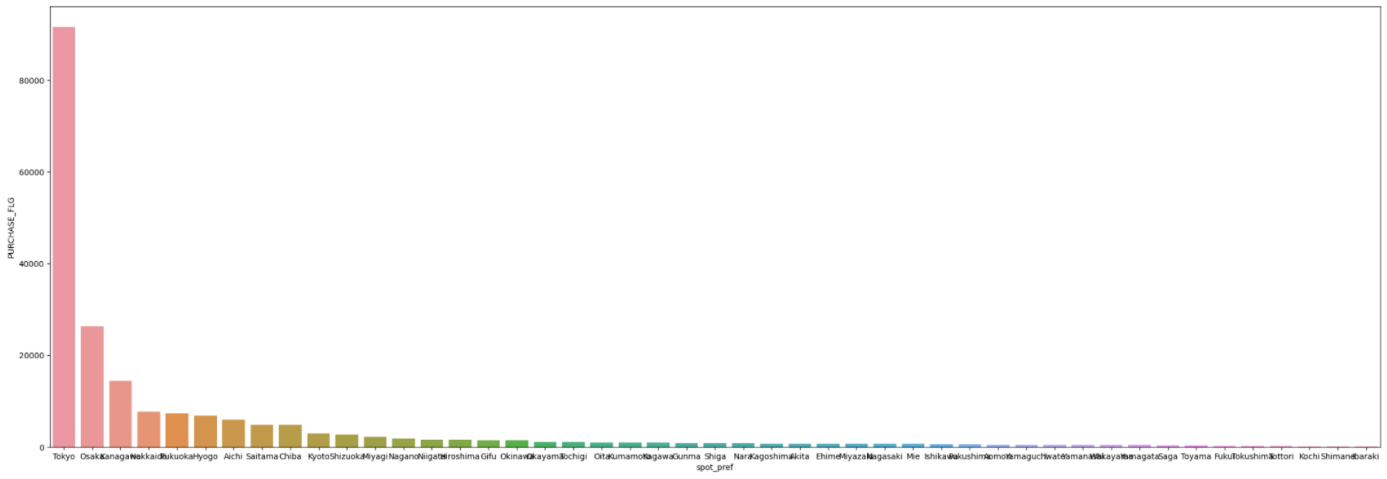
We can observe from the above Corelation plot that all features are significantly connected to one another. Price is substantially correlated with CatLog and Discount Prices, and Discount Prices are likewise correlated with CatLog Prices. As a result, we can conclude that Price is a significant characteristic of the dataset.



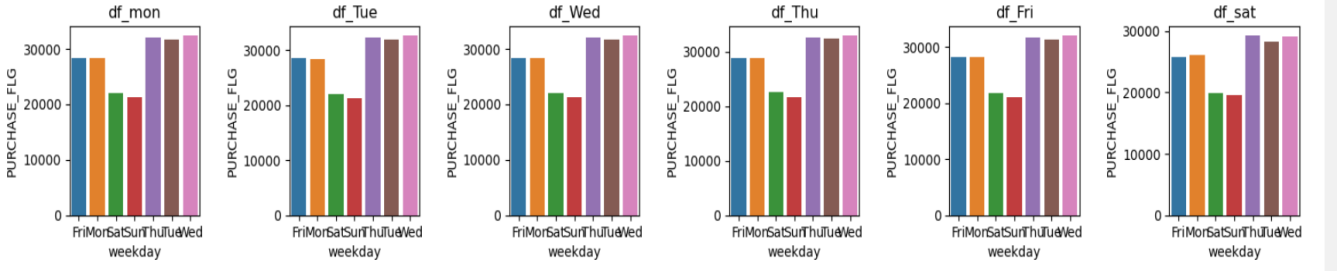
We can see from the above Bar Plot which case categories have the highest amount of Coupons used.



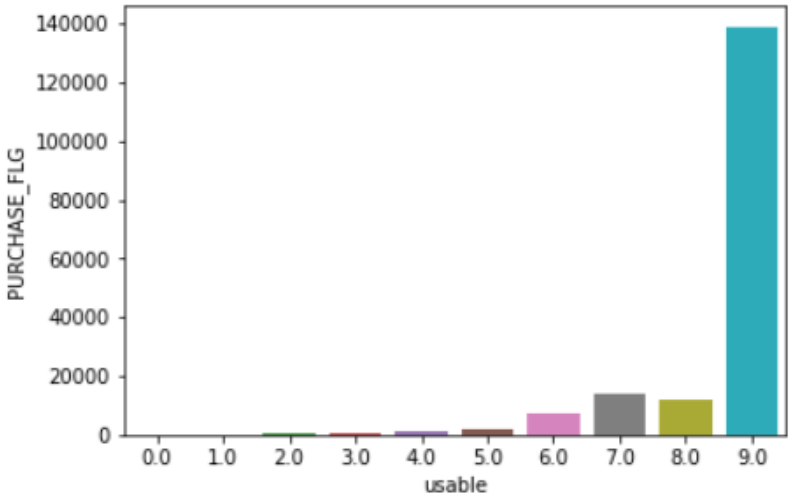
The Bar Plot above shows where the Coupons have been used the most.



The above graph depicts the utilisation of coupons per city, with Tokyo having the most coupon users, followed by Oskar.



We can see from the graph above which days of the week the customers used the coupons the most. Customers have utilised the most coupons for purchases on Tuesdays, Thursdays, and Wednesdays. Because the Ponpare Company offers large discounts on certain days, there is a significant volume of product purchases as a result, coupons are used the most on these days.



The greater the number of days that may be used, the higher the sales. The 9th of September was the last day for all food vouchers to be used.

There appears to be no specific product that sells on a particular day, and overall sales are comparable, with the third (Tuesday) through fifth (Thursday) days of the week having the highest sales volume.

Results:

Hypothesis Testing Results

(I) Independent Chi-Square test for User id and Coupon id

H0: Types of User id and Coupon id are unrelated (User id and coupon id have no relationship).

H1: User id kinds and Coupon id are not mutually exclusive (user id and coupon id have a relationship).

Chi-Square Analysis

The significance level is smaller than the P-value (2.2e-16) (0.05). As a result, the null hypothesis is rejected.

Model Evaluation and Performance

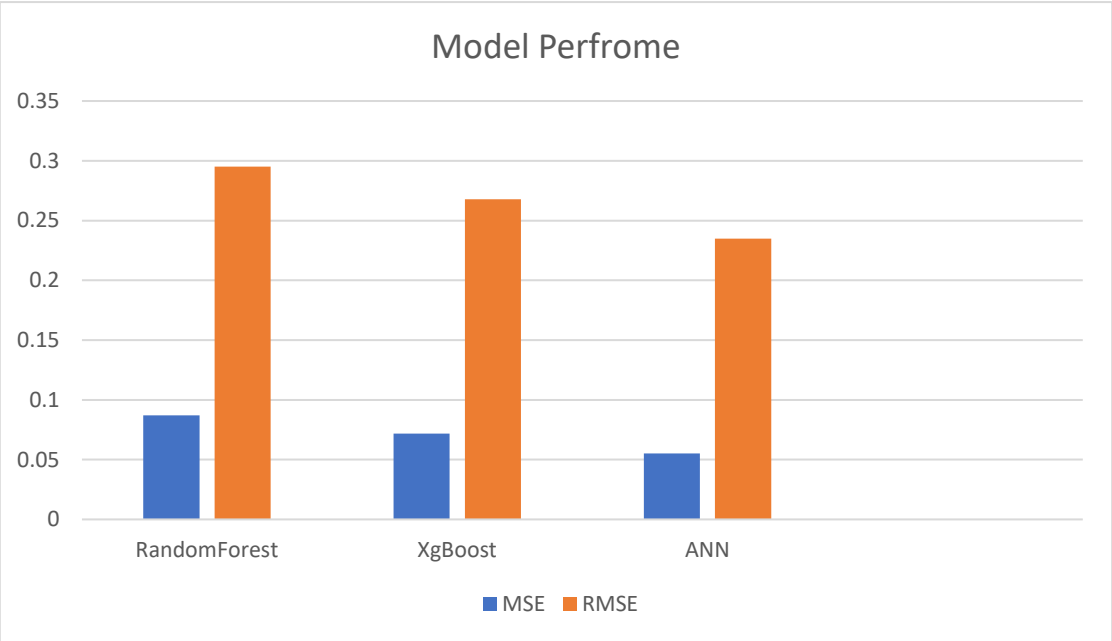
We divided our data into two subsets, with the training dataset accounting for 70% of the total and the test set accounting for the remaining 30%. USER ID, COUPON ID, PRICE RATE, and other parameters are supplied. We employed three different prediction models to train this dataset: the Rando Forest Regressor, XG Boost Regressor, and Artificial Neural Network (ANN). where we used a matrix evaluation method known as Root Mean Square Error to compare the model performance of three models.

When comparing these three models, we discovered that ANN performs better than the other two models we utilised.

The results are presented in a table format with three models below.

Sl.No	Models	MSE	RMSE
1	Random-Forest Model	0.08707	0.2950
2	XgBoost Model	0.07181	0.2679
3	ANN Model	0.05521	0.2349

The evaluate metric used was RMSE, with just the top coupon forecasts mattering for each user. The sequence in which the forecasts are made is also important. Our ANN model scored higher, as indicated in the table, with an RMSE of 0.2349 and a forecast of all 22875 participants or users.



Conclusion

The advantages of using the cosine similarity model are its speed and efficiency, low computing cost, and broad application: it scored all 310 coupons in the test dataset for each consumer. The disadvantage of my model is that I did not use all the data available.

Future Works:

For future study, we'd want to use either gradient-decent, LASSO, or other Deep Learning Models approaches to more precisely define the important weights. Second, we'd strive to include the impacts of VISITS and USER characteristics. Finally, we'd attempt a new model, such as a gradient boosted decision tree model, which produces high ratings for other competitors.

References:

https://www.projectpro.io/student-project/revanth-reddy-katanguri-coupon-purchase-prediction/16

https://www.galitshmueli.com/sites/galitshmueli.com/files/Team%209%20Tmall%20Loyal%20Customers.pdf

http://www.ijsrd.com/articles/IJSRDV7I21146.pdf

https://www.kaggle.com/c/coupon-purchase-prediction

https://rstudio-pubs-static.s3.amazonaws.com/136006_19010ce2f5144f4c89556a3eee159c57.html