Using Successful Franchise Data for Prediction of a Novel Cafe Location in New York City

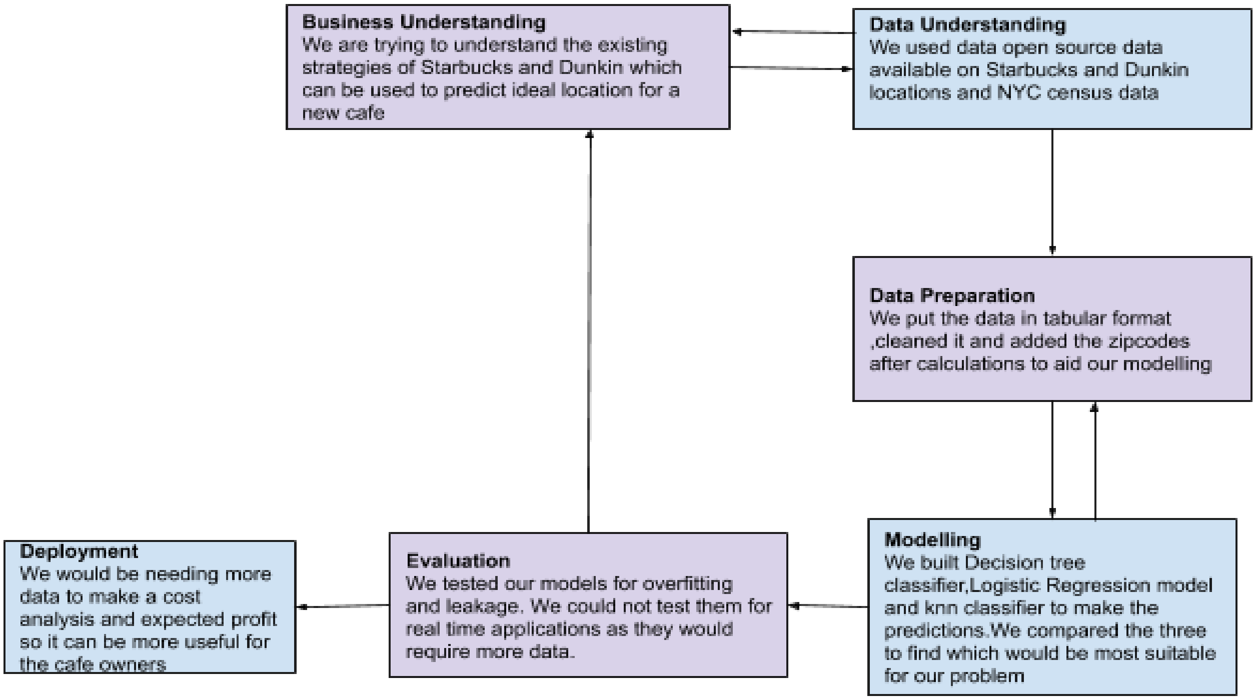
Abhijeet Sant, Nikita Amlani, Shreya
Stern School of Business, New York University

ABSTRACT

This paper is a definitive study to help upcoming café owners decide the location for their flagship store in New York City. We have designed a predictive model which studies the most important demographic factors, geographic coordinates and the existence of other departmental stores like Walmart and drug stores like Rite Aid that influence the location of Starbucks and Dunkin' Donuts to find their unique café strategies. A comprehensive consideration of New York City's census data along with different predictive models best suitable for this data driven decision making has been briefly discussed. The quantitative results have shown the presence of a Rite Aid store and Population Income (of that zone) as the most correlated factors in deciding a new flourishing coffeehouse. In consideration of the results obtained for NYC, the future scope of this project would be extending the model for the entire country.

1. INTRODUCTION

"A small café, that's love" enunciated Mahmoud Darwish. Coffee shops have become widely desired and extremely profitable not only for serving quality coffees, assorted pastries, and appetizing snacks but also for their relaxing aura and trending vibe. If starting from scratch, fresh Café chain Business owners have to invest a great deal of effort into the whole process and the key factor in determining the future, even before setting up the menu, would be to find a preferred and easily traceable spot for students to catch up on schoolwork, professionals for their unofficial business meetings and other ordinary humans to fraternize. Starbucks, the pioneer of coffee shops owns more than 31,000 coffeehouse locations around the world. No wonder new café owners have the pressure of getting it right in this competition-driven market where business owners are always looking for an investment opportunity at the right location. Research reveals that this decision making lacks the sufficient theoretical foundations and hypothesis, imperative to set in motion the wheel of Café business. This paper will guide novel café owners to open their first coffee shop in New York City by taking advantage of Data Science techniques given the required elements of considerations and countless deciding factors. In this paper, we detail an exceptional strategy of finding the right location for business owners in New York City with the help of wide datasets from Starbucks, Dunkin Donuts, and US Census data.



Business Understanding
We are trying to understand the existing strategies of Starbucks and Dunkin which can be used to predict ideal location for a new cafe

Data Understanding
We used data open source data available on Starbucks and Dunkin locations and NYC census data

Data Preparation
We put the data in tabular format ,cleaned it and added the zipcodes after calculations to aid our modelling

Modelling
We built Decision tree classifier,Logistic Regression model and knn classifier to make the predictions.We compared the three to find which would be most suitable for our problem

Evaluation
We tested our models for overfitting and leakage. We could not test them for real time applications as they would require more data.

Deployment
We would be needing more data to make a cost analysis and expected profit so it can be more useful for the cafe owners

Figure 1. The CRISP Data mining Process for our project[6]

2. BUSINESS UNDERSTANDING

Businesses are increasingly driven by data analytics and by great professional advantage of being able to interact competently with and within such businesses.Understanding the fundamental concepts and having frameworks for organizing the data analytics thinking allows us to interact competently and envision opportunities for improving data-driven decision-making.

The project deals with decisions for which ‘discoveries’ need to be made within the data and so decision-making can benefit from even small variations in decision making accuracy based on data analysis.

The aim of this project is to find the ideal location for the setup of a new cafe in New York City.To achieve this understanding of the existing business strategies of the two popular cafes, Starbucks and Dunkin donuts can be helpful. Data from store locations and the demographic details of these neighborhoods will assist us to predict the locations of these stores. By doing so we can expect that if we open our cafe in a location based on similar features it can be successful.

Multiple supervised learning classification models for Starbucks and Dunkin with target variables as Is_Starbucks and Is_Dunkin which would take binary values are built.We carefully studied the features which are contributing to the prediction so as to ensure there was no leakage and overfitting. Decision Tree , Logistic Regression and k-NN models were built to compare the performances simultaneously using training data and then tested the models on testing data. Using this business proposal cafe owners can understand which features of a location are taken into consideration while setting up these cafes and can also choose a similar location for higher profit and favourable outcome.

3. TECHNICAL PRELIMINARIES

For solving the current business problem, the top 2 coffeehouse corporations of the USA, Starbucks and Dunkin are being considered. An in depth reflection on their strategy of success, about how they decide on location selection, what can their store locations tell us about their customer base and what they view as attractive locations to serve this base has been made. Can we infer at a high level where they should or will explore further store expansion? The analysis kicks start by grounding ourselves in an overall view of stores in NY. A total 645 stores for Starbuck (until 2017) and 1356 stores for Dunkin (until 2019) were reported. Datasets were available over Kaggle[1]

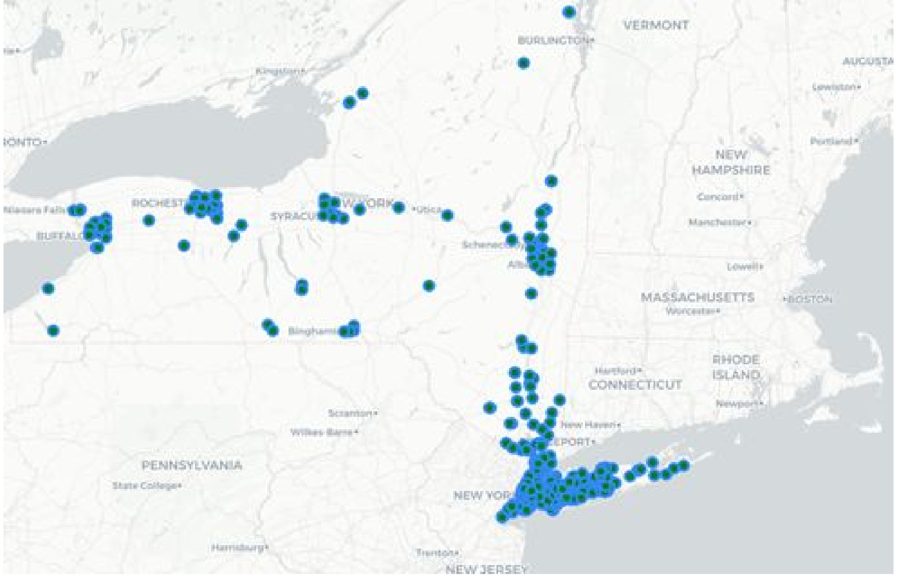


Figure 3.1. Visualization depicts Starbucks location and darker regions means more stores. Data is as of 2017 and will not include more recent openings.

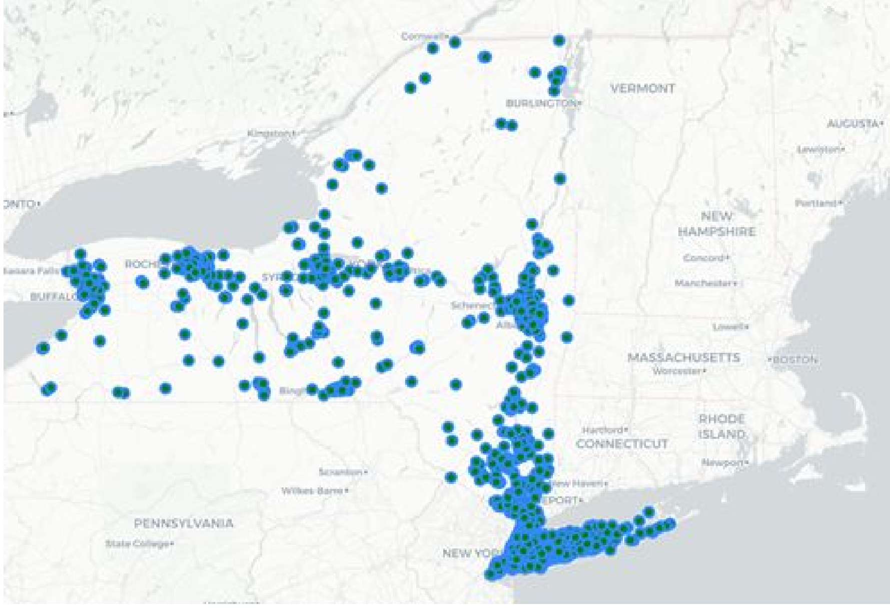


Figure 3.2 Visualization depicts Dunkin 'location and darker regions means more stores. Data is as of 2019 and will not include more recent openings.

The above visual image depicts a bias toward the coasts and major metropolitan areas, reasonable given the context of being an upscale mass-market brand seeking consumers that can pay slightly premium prices.

4. DATA UNDERSTANDING

To get a better idea about their customer base, a dataframe was created with Census data available over Kaggle[2]. Features like total population, income per capita and other demographics of all the locations were investigated. The census data provides a better understanding of what part of one of the largely populated cities of the US has the highest number of Franchise locations per capita. Considering NY, one of the densely populated states in the US, having a roughly population of 19.45 Million, we can estimate around 30155 people per Starbucks and 14343 people per Dunkin' location. The scope of this project is limited only to NYC. By analyzing store data and how that maps to Census demographics, a high-level "profile" of what Starbucks has deemed to be a valuable location, in addition to seeing where there may be opportunity to further expand footprint in and around these areas can be created. Below is a box-plot of the population and income per Capita with respect to the presence of Starbucks or Dunkin.

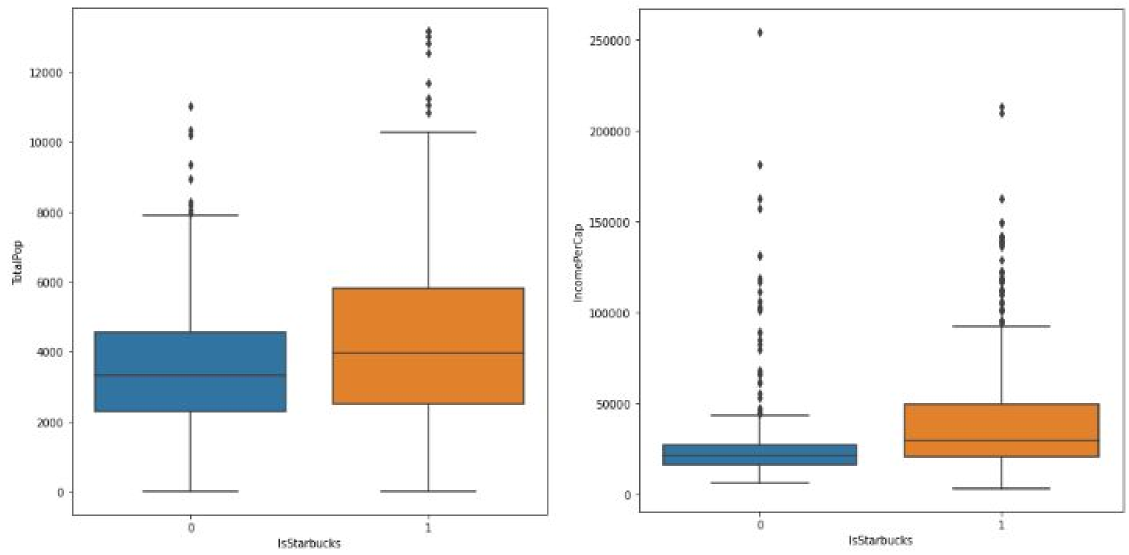


Figure 4.1 Box plot representation of Total Population and IncomePerCapita of Starbucks dataset.

In Figure 4.1, the plot makes clear that it targets locations with wealthier households (Income Per Capita of ~$165K vs. $150K), lesser people (population of ~21.5K vs.~2K). The wealthier locality (customers), the more restaurant locations.

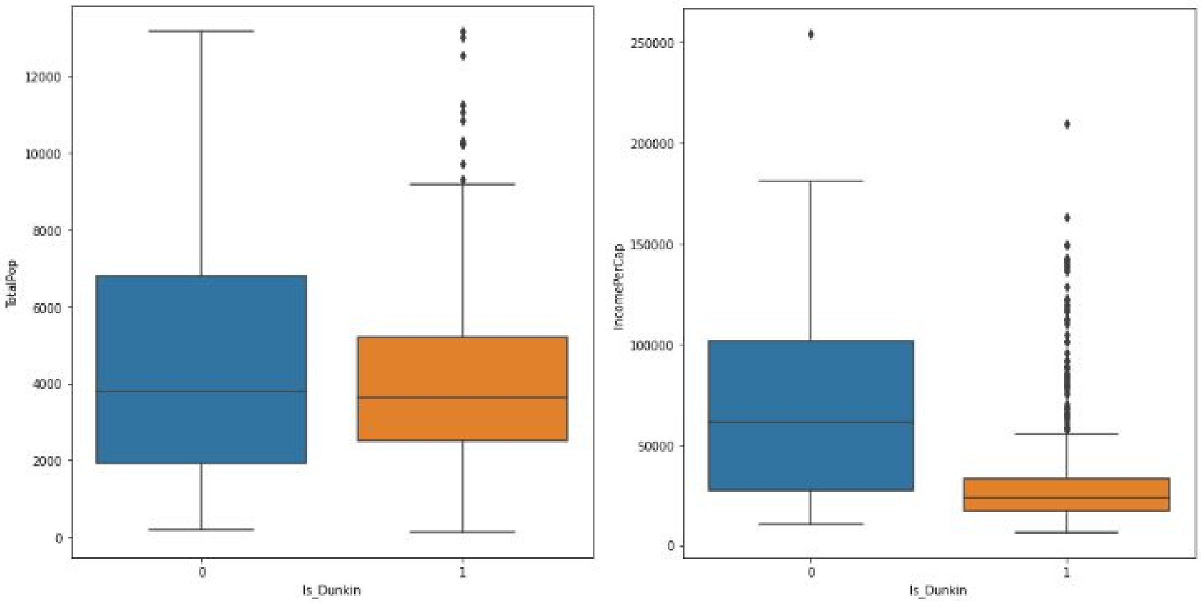
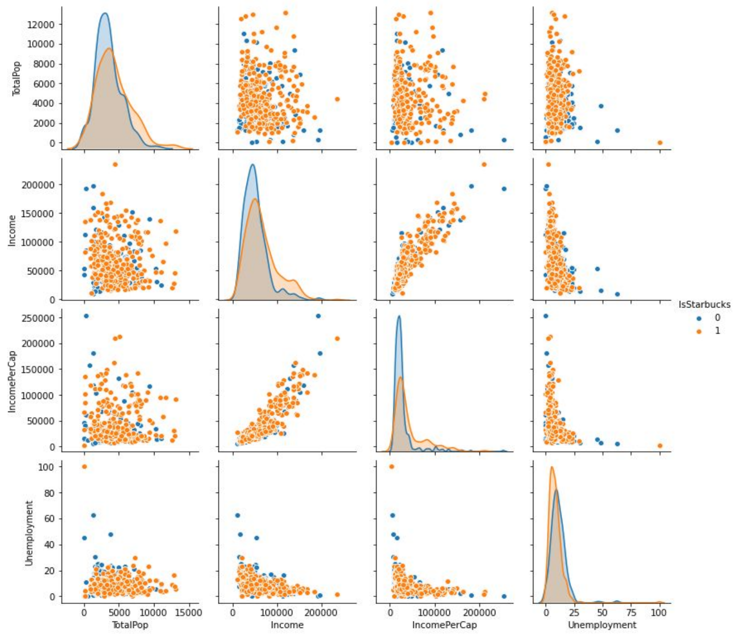


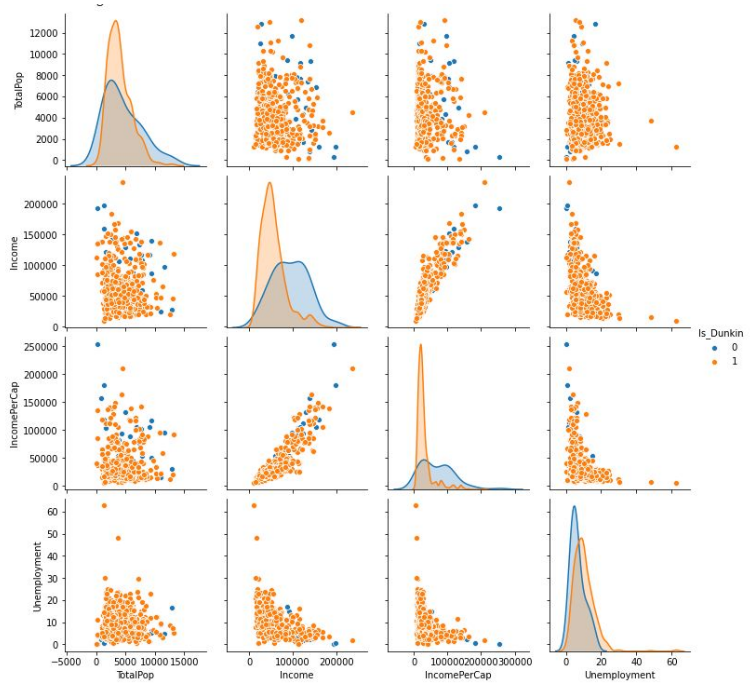
Figure 4.2 Box plot representation of Total Population and IncomePerCapita of Dunkin dataset.

In Figure 4.2, the plot makes clear that it targets locations with lesser wealthier households (Income Per Capita of ~$165K vs. $350K), but more people (population of ~3K vs.~3.2K). The more people (or households), the more restaurant locations.

A PairPlot is highly effective to understand multivariate relationships according to the target variable (‘Is_Dunkin’ or ‘Is_Starbucks’). It shows every variable’s association with respect to one another and each variable’s marginal distribution along the diagonal.



Plot depicts how well some of the feature variables split the data according to the ‘Is_Starbucks’ variable. Variables depict that Income Per Capita (IncomePerCap) gives the tightest and most correlations with Is_Starbucks. Income and IncomePerCap are highly correlated. By losing one of the variables, we can reduce the dimensionality.



Plot depicts how well some of the feature variables split the data according to the ‘Is_Dunkin’ variable. Variables depict that Total Population (TotalPop) gives the tightest and most correlations with Is_Dunkin. Observe that Income and IncomePerCap are highly correlated. By losing one of the variables, we can reduce the dimensionality.

5. DATA PREPARATION

The collected data consisted of Starbucks' locations, Dunkins' locations and NY census data. The data set was not exactly aligned for the problem statement, the only advantage being geo-location of each store and census data had satisfactory income related information for that location as these data sets were available virtually for free. Each of the data sets were converted into tabular format i.e. a DataFrame and the rows dropped which had missing values in any of the columns using DataFrame.dropna() using Pandas in Python. Data mining technique requires both categorical and continuous data, hence converting those data to a different format was essential. The numerical values were scaled for better comparability while modeling.

As the data received from Kaggle was clean enough, we validated its relevance by analyzing its describe() and corr() function. The variables of interest for this blueprint were in a numeric format; hence transforming them for type casting was not necessary. With the help of the Latitude and Longitude variable data, ZipCode of each row was calculated. Using USZipCode.SearchEngine() function to create a new column of ZipCode, as this column is common in each of the 3 data sets, we decided to use it to clean and match customer records and ensure only one record per customer existed.

Sorting each row with respect to each of the columns, and deleting duplicates was done under the cleaning process. This indirectly removed data leakage (otherwise, we might have ended up with a few common rows in Training and Test data). Next we checked all the possibilities of leakage in our data sets. As we know, when a training set has a dependency on the test set, this might also lead to data leakage; but this was not the situation in our case. Secondly, the models were fit with the training set and evaluated for predictions on the test set. This helped in achieving a precise accuracy.

Starbucks and Dunkin not only has more locations in proportion with population growth (i.e. higher population states have more stores) and IncomePerCapita, but also have over index in states with major metropolitan areas (i.e. these locations have more stores overall and more stores per person). Data sets for Pharmacies, Corporate Retail shops or Apparel retail shops were simultaneously explored. In our understanding, we were specifically looking for CVS, Walmart, Zara, H&M locations. Among these, datasets were available for Walgreens, CVS, Walmart, Rite Aid, Super Stop n Shop which took some effort to obtain. Unfortunately, data sets for Zara and H&M didn't not exist. Attribute selection for feature engineering was done, after estimating the cost and benefit of each data source.

Finally, a DataFrame which was the combination of the Census data, a few binary columns i.e. Is_Starbucks, Is_Dunkin, Is_Walgreens, Is_CVS, Is_Walmart, Is_RiteAid, Is_StopnShop was modeled. These columns shed light on the idea of ZipCode also consisting of these retail shops. Is_Dunkin was a target variable for Dunkin donuts with the understanding of creating a model which could predict if the

ZipCode has a Dunkin store in its area or not.Similarly Is_Starbucks was a target variable for Starbucks modelling.

6. MODELLING

Models in data science can be segmented into supervised and unsupervised learning algorithms. This project deals with supervised learning models(k Nearest Neighbours, Logistic Regression and Decision Tree) since we are considering labelled datasets and a target variable. This phase consists of the below points and demonstrates a generalized approach.

a. Defining Predictor and Target columns

While exploring and understanding the data phase, the dataset is defined for predictor and target columns. The attributes that aids in prediction are referred to as predictor columns and the target column is defined as the attribute of a supervised model that is tried to predict.

After performing feature selection, the following are the narrowed predictor attributes:

Dataset Group	Predictor Attributes
Census Data for NY	State, County, Borough, TotalPop, Men, Women, Hispanic, White, Black, Native, Asian, Pacific, Citizen, Income, IncomeErr, IncomePerCap, IncomePerCapErr, Poverty, ChildPoverty, Professional, Service, Office, Construction, Production, Drive, Carpool, Transit, Walk, OtherTransp, WorkAtHome, MeanCommute, Employed, PrivateWork, PublicWork, SelfEmployed, FamilyWork, Unemployment
Starbucks/Dunkin Dataset	Latitude, Longitude, ZipCode
Other Stores’ Location Dataset	IsWalmart, IsWalgreen, IsCVS, IsRiteAid, IsTarget, IsSupershop

The considered target attribute are mentioned below:

Dataset Group	Target Attributes
Starbucks/Dunkin Dataget	Is_Starbucks/Is_Dunkin

b. Splitting into training and testing data

The standard train and test split to sample a general dataset is 80% training data and 20% testing data.

c. Fitting the model

A model can be loosely defined as a mathematical description or a statistical plan. Fitting this mathematical representation in any dataset is one of the primary goals for a data science project. Implementation of three modelling techniques for this project can be tracked below.

d. Obtaining Prediction from the model

Prediction of the target attribute by analysing the characteristics of the dataset is the central theme.

e. Evaluating Performance of the model

When determining the perfect model, statistical values like CV, ROC-AUC and accuracy are measured and compared for the model predictions on the test data.

The above process is re-visited to get the optimum complexity parameter and better accuracy and applied to the below three in-scope supervised machine learning models:

a. Model 1: k-Nearest Neighbours

This similarity based model works by finding the distances between the query data point/instance and all the examples in the data by choosing the specified number of k points adjoining to the query instance and then by voting the most frequent label or average of the label. The complexity parameter in this model is k, as the value of k increases, predictions become more accurate whereas after decreasing the value of k, predictions become less stable. After having sought multiple values, k=10 was seen as an ideal fit.

b. Model 2: Logistic Regression

This linear based classification model uses logistic functions like Sigmoid for predictions. This predictive analysis technique requires no high correlations and no outliers among the predictor attributes. To assess this, a correlation matrix among the predictors can be used. Tabachnick and Fidell(2013) Correlation coefficients for independent variables less than 0.90 meets the requirements.

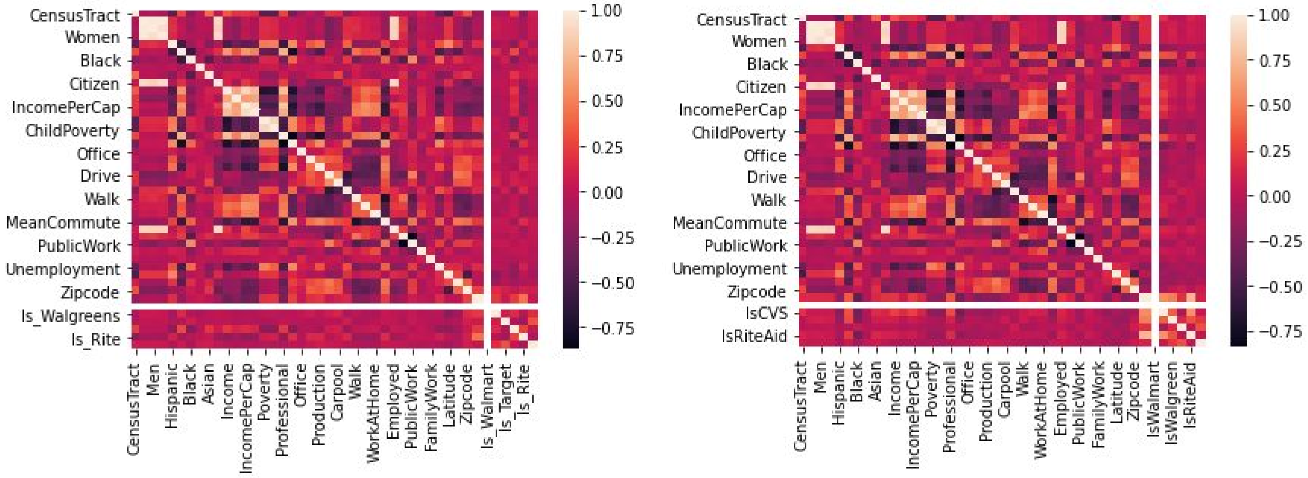


Figure 6.1 Heatmap of Dunkin Dataset (on the left) and Starbucks (on the right) to show correlation between attributes

To avoid leakage, SHAP value calculation was performed. Detailed SHAP values i.e impact on the model output values along with their feature vectors can be found below:

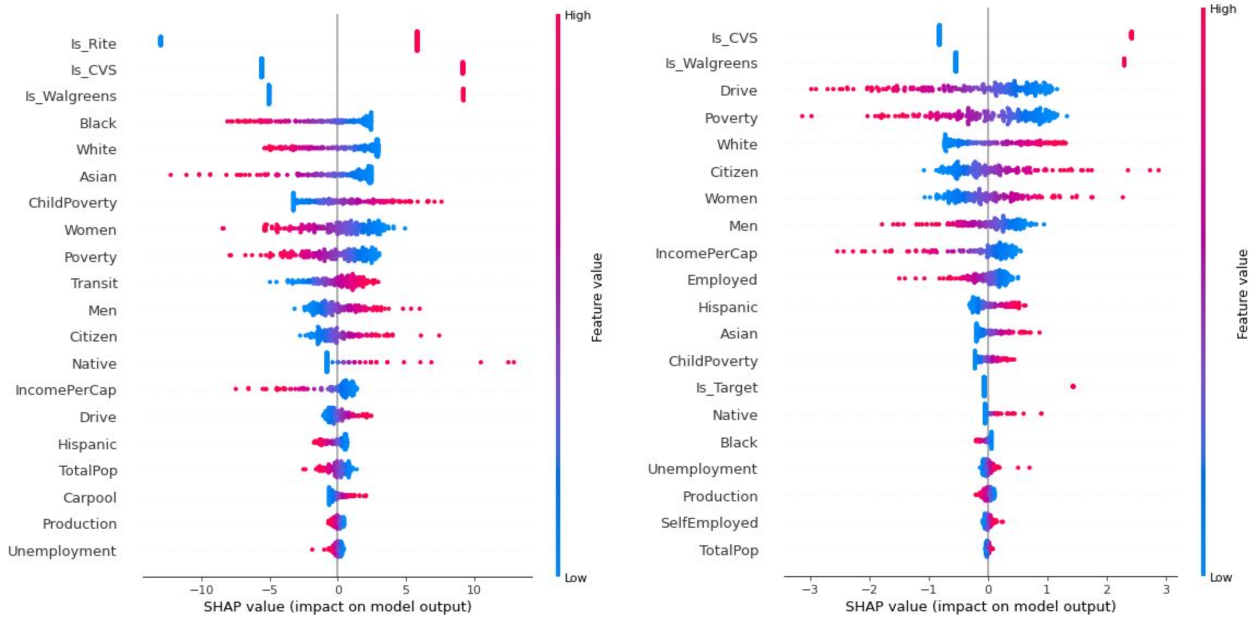


Figure 6.3 SHAP values against features for Dunkin and Starbucks are presented above respectively

c. Model 3: Decision Tree

This tree based model contains nodes that are conditioned over a feature. This condition is the decisive factor for the nodes to navigate and assemble an inverted tree, branched off from a homogeneous root node to highly heterogeneous leaf nodes. For selecting a proper condition and making the tree more efficient Information gain and Entropy has been used. Entropy is a measure of disorder (calculated on the Target Column) and Information gain is the change in Entropy. The highest Information gain was obtained on Is_Rite and IncomePerCap from US Census dataset for both i.e. Starbucks and Dunkin Donuts. Following are the equations for Information gain and Entropy:

$Entropy = - \sum p(X) \log p(X)$

$Information\ Gain\ (Parent,\ Children) = Entropy(Parent) - Entropy(Children) * Weighted\ average$

Decision trees are highly prone to complexity and overfitting. To handle such concerns, calculation of accuracy over multiple tree depths was performed and max_depth was set to 5, making it most precise for both the datasets (Dunkin and Starbucks).

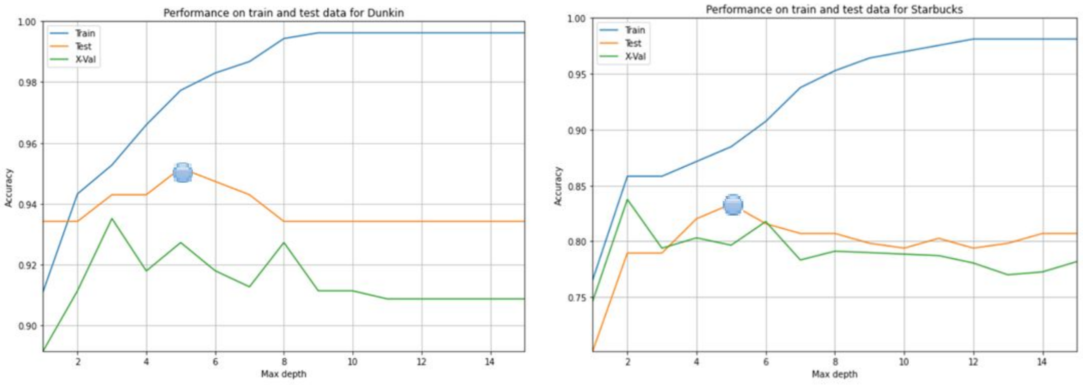


Figure 6.4 Accuracy over training and testing data can be viewed and the blue point shows the chosen Max-Depth on both the datasets.

For better comprehension, review the below Decision Trees for Dunkin Donuts and Starbucks.

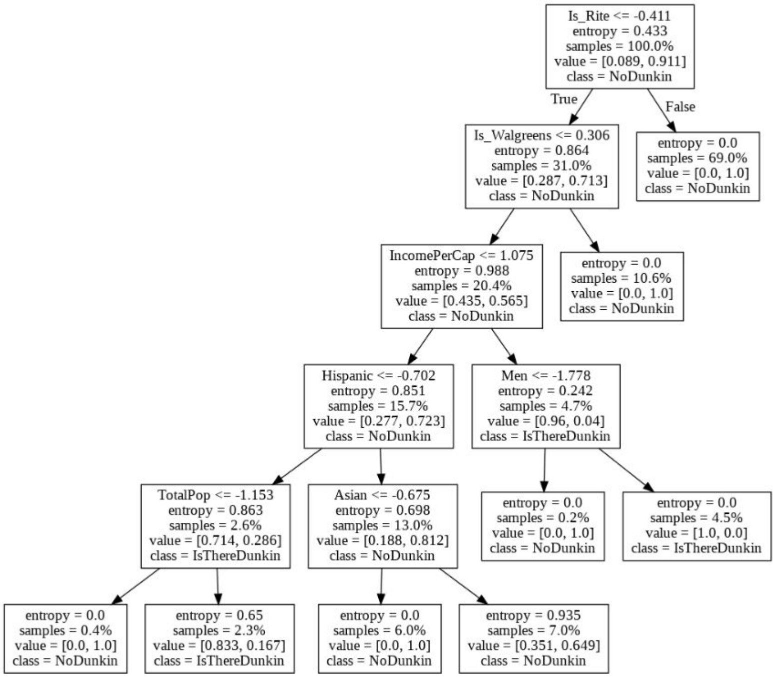


Figure 6.5 Decision Tree embedded with Information Gain and Entropy statistics for Dunkin Donuts

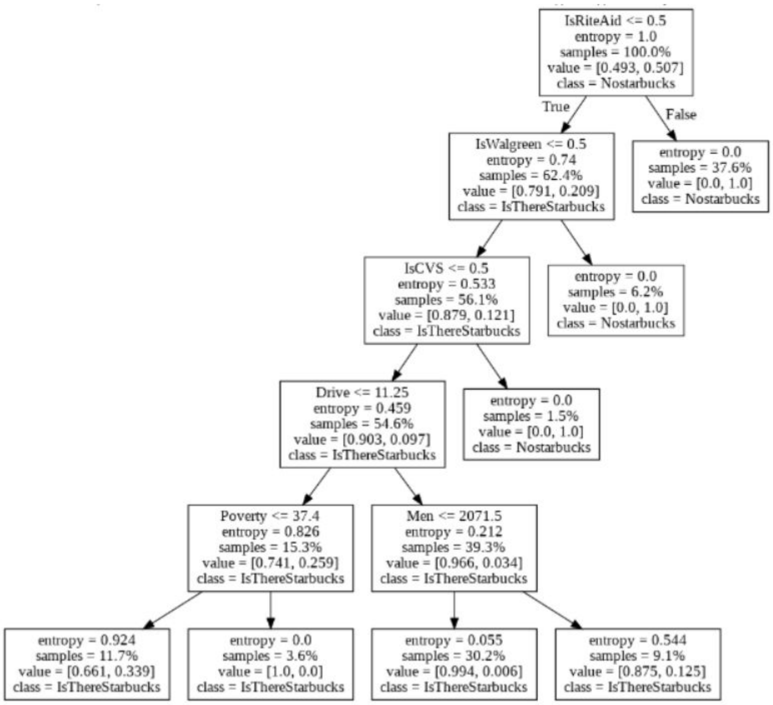


Figure 6.6 Decision Tree embedded with Information Gain and Entropy statistics for Starbucks

Calculation of SHAP values to look for leakage while modelling the decision tree was carried out. The results have been discussed in the Evaluation phase.

Model Comparison and Selection

Following advantages of Decision-Tree over k-NN and Logistic Regression were witnessed:

a. ROC-AUC measure of Decision-Tree was found to be more stable.

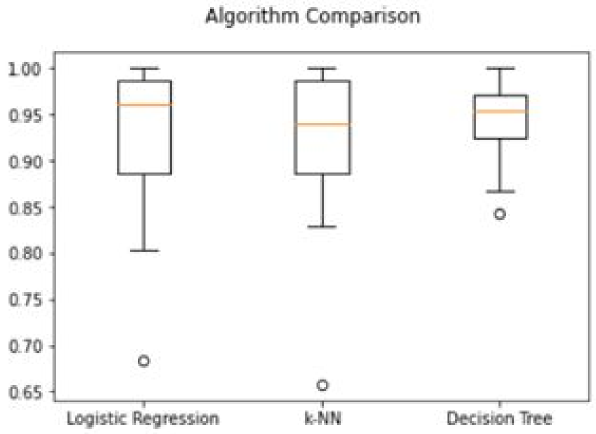
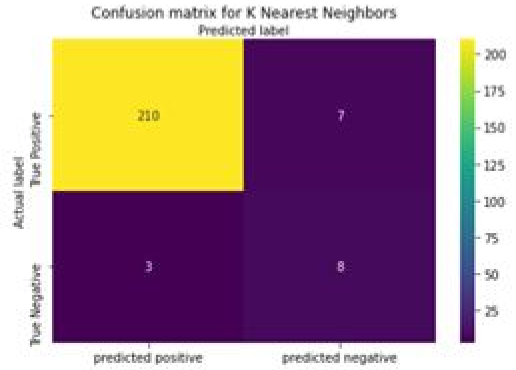
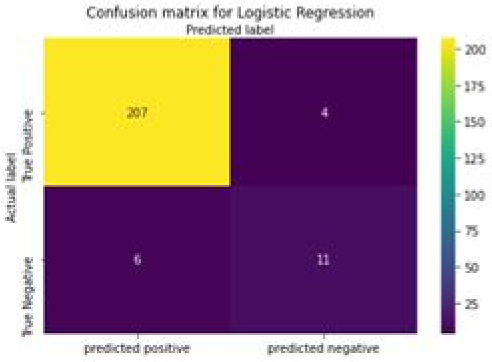
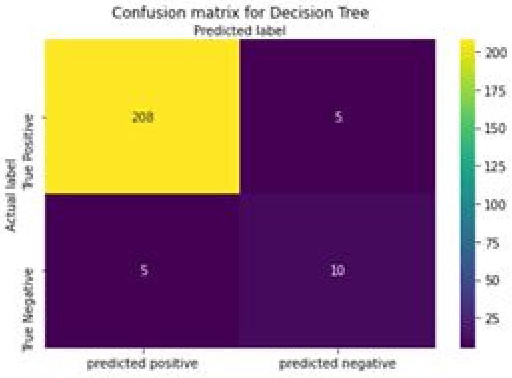


Figure 6.7 Box-Whisker plot to compare ROC-AUC accuracies of the above three supervised models

b. Confusion Matrix of Decision Tree vs the others for Dunkin Donuts dataset:







Similar results were obtained for Starbucks dataset as well.

c. Decision Tree was comparatively faster and provided an understandable explanation by examining the IG and Entropy values.

Hence the Decision Tree model is considered the most appropriate choice for this business proposal.

7. EVALUATION

For this project, multiple modelling techniques were carried out to figure out the best business strategy referencing Starbucks and Dunkin. Logistic Regression and Decision Tree Classifier gave us a clear picture of the unique strategies of placement of these cafes. To ensure that these models are working correctly some evaluation techniques were examined closely.

1. **Accuracy Of Models -**

For Starbucks -

a) Logistic Regression -
This model gave an accuracy of 96.03 % on training data and 95.18% on Test data.

b) Decision Tree -
This model gave an accuracy of 95.46% on training data and 92.54% on test data.

For Dunkin -

a) Logistic Regression -
This model gave an accuracy of 96.79 % on training data and 95.61% on Test data.

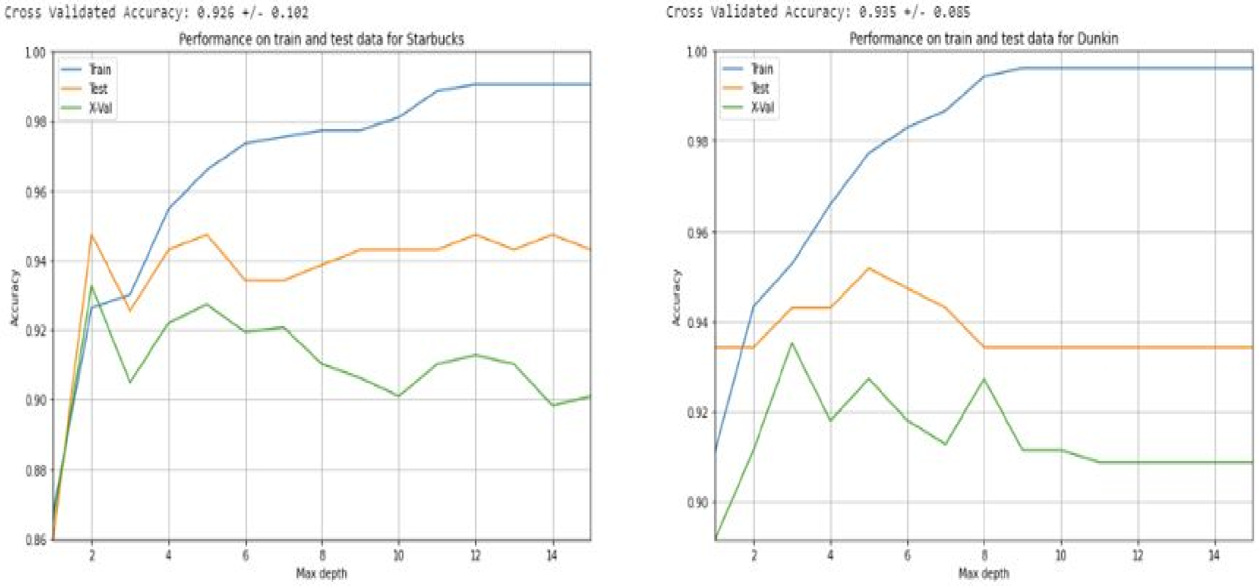
b) Decision Tree -
This model gave an accuracy of 97.61% on training data and 95.61% on test data

2. Base Rate -

Calculation of the base rate for both the stores was necessary for collating other accuracy based metrics. The Starbucks dataset provided a base rate of 50.46% while Dunkin had a base rate of 91.3% .

3. Cross-Validation -

The next step was to use cross-validation to check if the higher accuracy was due to overfitting .
For Starbucks the cross val score for Decision tree classification was 92.6 with standard deviation 0.102 and for Dunkin it was 93.5 with standard deviation 0.040 . The standard deviation is low that indicates there is very low deviation between folds.



```
Cross Validated Accuracy: 0.926 +/- 0.102
```

```
Cross Validated Accuracy: 0.935 +/- 0.085
```

Figure 7.1 Cross validation curves for Starbucks and Dunkin for different models

4. Plotting ROC Curves -

We then plotted ROC Curves for both the models . The AUC score is closer to 1 indicating that the models are able to predict correctly.

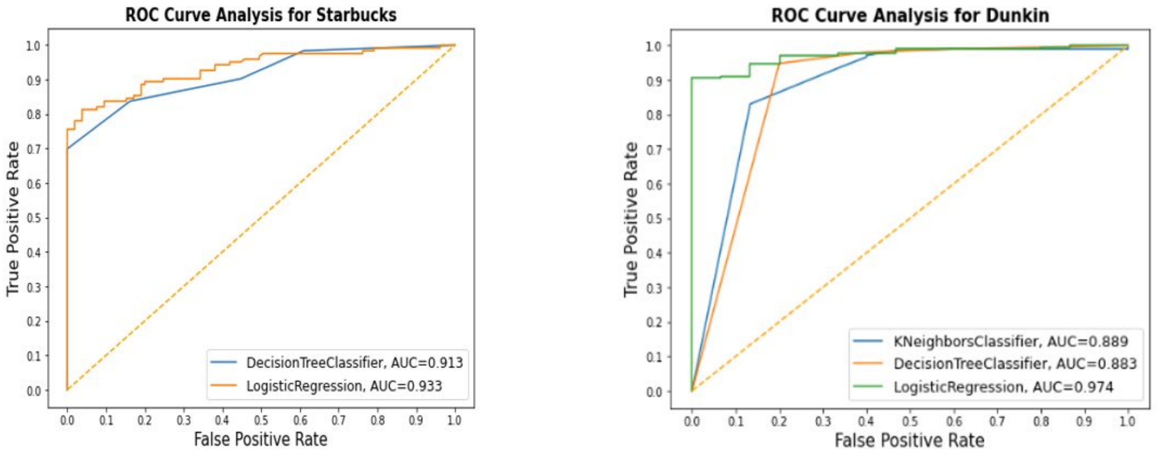


Figure 7.2 ROC curves for Starbucks and Dunkin for different models

5. SHAP Analysis -

To find the contribution of different features and find leakage for any features, SHAP analysis was used. SHAP values for each feature represent the change in the expected model prediction when conditioning on that feature. For each feature, SHAP value explains the contribution to explain the difference between the average model prediction and the actual prediction of the instance. Below we have plotted a variable importance plot that lists the most significant variables in descending order. The top variables contribute more to the model than the bottom ones and thus have high predictive power. An observation was made that having retailers and pharmacies around will highly affect the predicted location of cafes . There are some features which contribute separately to predicting where there is a Dunkin or Starbucks in that location which can be seen from the SHAP analysis plot below. The details of the strategy used by the stores individually are explored in the section below.

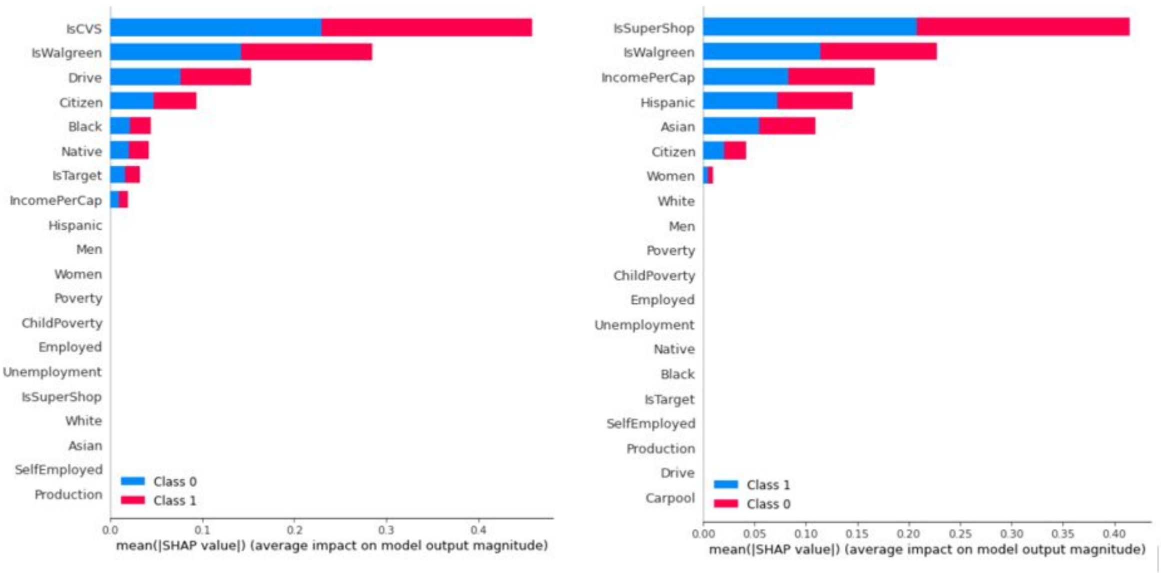
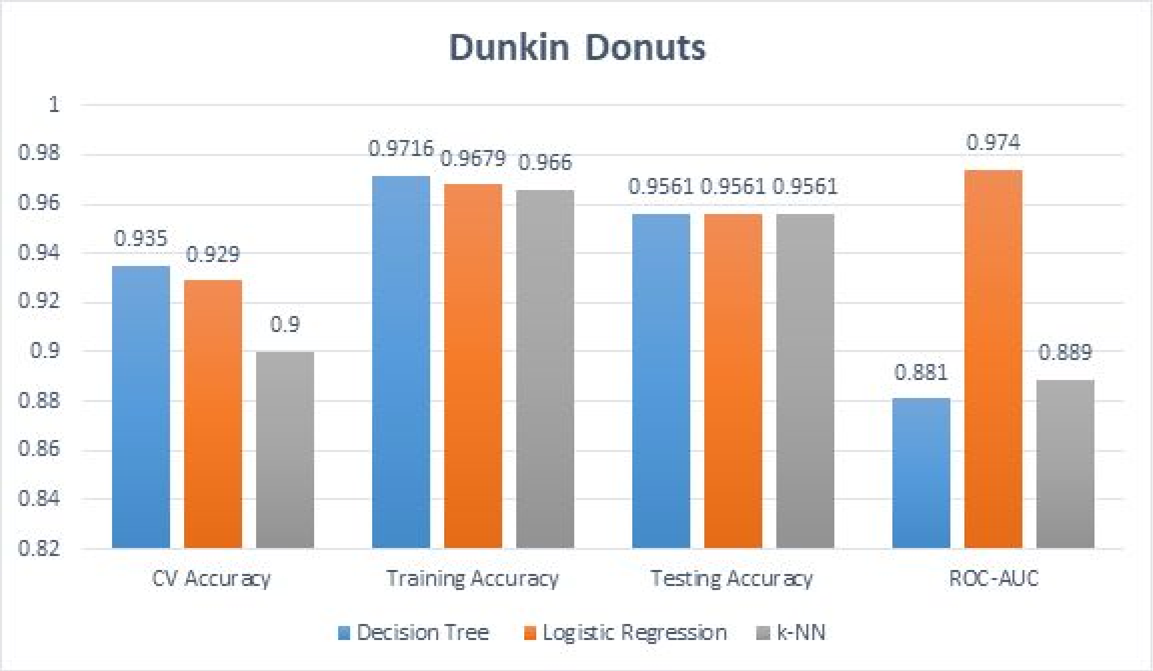


Figure 7.3 SHAP Analysis for Starbucks and Dunkin for Decision Tree

6. <u>Model Performance Evaluation Summary</u>



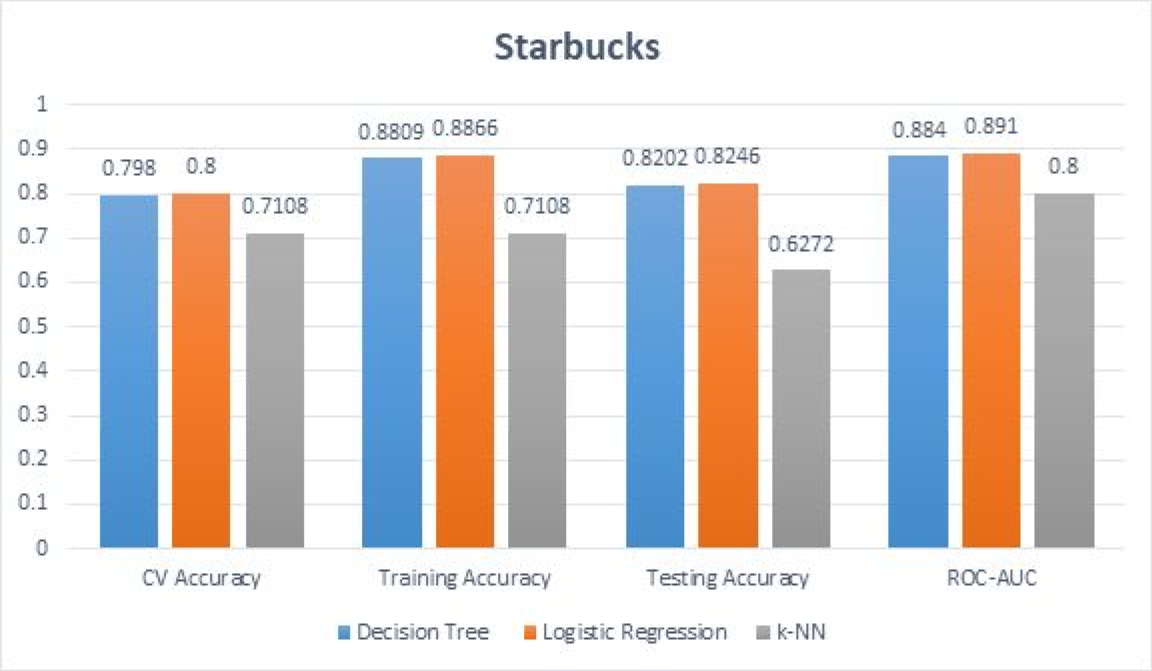


Figure 7.4 Performance plots of all three models

8. STRATEGIES DETAILS USED BY STARBUCKS AND DUNKIN

After evaluating the models and testing if they are overfitting or not, information regarding business strategies used by the two leading cafes was seen.

As of November 2019, Starbucks is present on 6 continents and in 78 countries and territories, with around 31,256 Locations[4].

Taking insights from data, the following strategy used by them to open stores -

1) They tend to be around big pharmacies and stores. This indicates that they target places where people tend to visit more.

2) There are stores where there is a more employed population, number of people using cars and higher per capita income.

Dunkin has 11,300 restaurants worldwide –that's over 8,500 restaurants in 41 states across the U.S.A. and over 3,200 international restaurants across 36 countries[5].

Dunkin is one of the affordable cafe which also serves some amazing donuts and munchkins which makes it a widely loved cafe, usually opens stores based on the following -

1) They also tend to be around big pharmacies and stores. This indicates that they target places where people tend to visit more.
2) There are more stores where there is more poverty and public transit is being used.

9. DEPLOYMENT

Deployment of the model simply means the integration of the data mining system into an existing production environment which can take in an input and return an output that can be used in making practical business decisions. Instead of deploying the whole model, going ahead with only the data mining phase as it will adapt with the changing future analysis techniques can incorporate with the location of new stores.

Deployment of this model meets the business idea of predicting if a location (ZipCode) is ideal for opening a new Café and having met the strategies used by top 2 coffeehouses cooperation of US, Starbucks and Dunkin.

Deploying the model into the production typically requires the model be re-coded from the production environment, usually from greater speed or compatibility with an existing system. This may incur substantial expense and investments. Hence before this work gets deployed, a working prototype can be identified and contrasted along with the leadership/marketing team for consultation with their business understandings. A proper scrutinization of the model from the domain experts or stakeholders before deployment is suggested. This would help in gaining more insights with development, procuring constructive feedback and working on them to improve this proposal.

Once a better understanding of the cost estimates is obtained from our business analyst, \the cost/benefit matrix that can be incurred for our model. This can assist in retrieving the expected value and the expected profit with respect to the location chosen for the new predicted Café Location. This version of the working prototype can actually best address the stated business problem. Only when a model is fully integrated with the business systems, it can be extracted for real value from its predictions.

10. CONCLUSION AND FUTURE WORK

To find the ideal location for prospective cafe owners to open their cafes so they can be successful was the central idea of this business proposal. To do so, the business strategy, census data, locations of various retailers and pharmacies and the store locations of the cafes in New York City used by two popular cafes Starbucks and Dunkin was contrasted effectively. Models which could predict if the location would have a Starbucks or a Dunkin and by doing examine the strategy they used to open a new store. After the conducted research some great insights about the two models were revealed. Someone looking to open a new cafe can use this model and find various locations where a cafe could be successful.

The above predictions and business proposal is for New York city, and can be considered as a smaller dataset(collation with the entire USA data) to evaluate and get information from. A pertinent strategy will be to expand this project to the whole of the US so that a generalized model can predict locations all across the country for owners all over. The data from profits and earnings of the cafes was untraceable, due to this predicting expected profits for different locations was out of the scope for this project however the pitch for a business owner would include the possibility of gathering this data for future use. Costs required to build the cafe in the location, how much the rent would be demanded and other miscellaneous costs were off the grid. These would also be needed to determine the holistic success of a cafe location. Investments must be made to acquire a diverse range of datasets to broaden our findings but the above findings provide an excellent blueprint to get the ball rolling for a niche cafe owner with limited data science experience.

References :

[1] https://www.kaggle.com/starbucks/store-locations

https://raw.githubusercontent.com/trendct/dunkin-donuts-ct/master/dunkindonuts.csv

[2] https://www.kaggle.com/muonneutrino/new-york-city-census-data#census_block_loc.csv

[3] https://towardsdatascience.com/shap-a-reliable-way-to-analyze-your-model-interpretability-874294d30af6]

[4] "Number of Starbucks stores worldwide 2018 | Statista". Statista. Retrieved June 20, 2018.

[5] https://www.dunkindonuts.com/en/about/about-us

[6] Data Science for Business by Foster Provost & Tom Fawcett

Contributions of each group member -

All three of us divided the work to be done and worked diligently to share our final results with all.
Abhijeet worked on data preparation and understanding. He identified various sources from where we could get the data and how we could use it. He then cleaned the data and prepared it so we could use it for our modelling.
Shreya worked on creating supervised modelling for Dunkin and its evaluation.
Nikita worked on creating supervised modelling for Starbucks and its evaluation.
We all worked on checking if the models were overfitting or if there was any leakage. We also divided the final report writeup and all of us wrote a part of it.