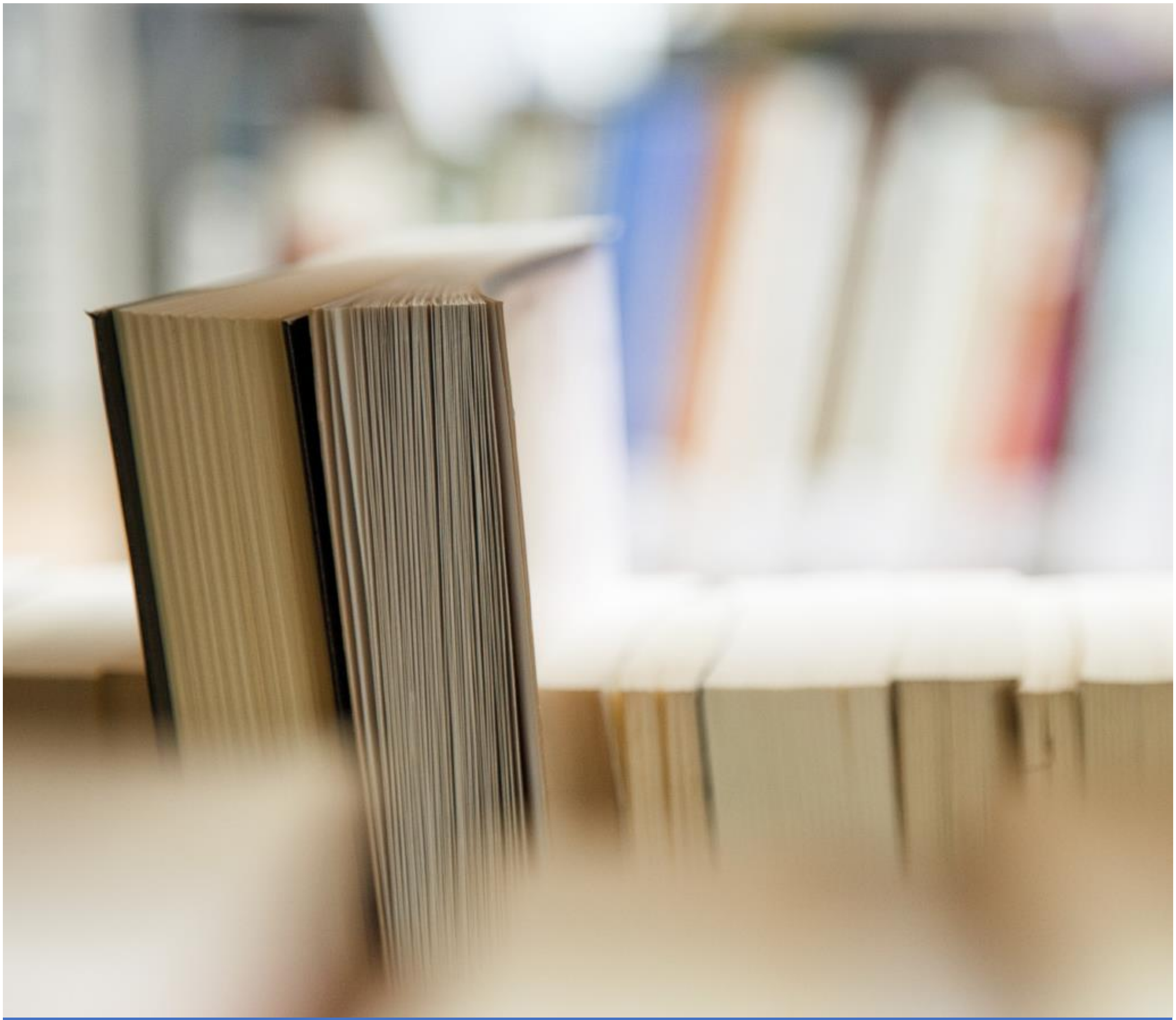


Springboard Data Science Career Track Capstone 1 – Milestone report

Walmart Trip Type Classification

Rajesh Dharmarajan

8/16/17

Springboard Career Track

TABLE OF FIGURES

1. OVERVIEW

As my Capstone project, I will be working on Trip Type Classification challenge from Walmart posted on kaggle.com. The details of this competition can be found at https://www.kaggle.com/c/walmart-recruiting-trip-type-classification

This document provides a summary of the problem, the approach taken to arrive at the solution.

1.1. PROBLEM STATEMENT

Walmart is a leading department store chain operating in multiple countries that has seen exponential growth to almost half a trillion-dollar enterprise in just last 55 years from its inception. As per the company, "Each week, over 260 million customers and members visit our 11,695 stores under 59 banners in 28 countries and e-commerce websites in 11 countries. With fiscal year 2017 revenue of $485.9 billion, Walmart employs approximately 2.3 million associates worldwide."

Walmart has been able to grow at such a rate by accurately ensuring that it provides exactly what the customer if looking for and by being able to predict through art and science what the customer would be looking for. One of the tools in Walmart's quest for better merchandising and customer service is the data about its customers purchase history. Walmart classifies each visit by a customer makes to its store based on the items she purchases in to different 'Trip Types'. This enables Walmart to create the best shopping experience for every customer.

1.2. OBJECTIVE AND EXPECTED RESULTS

Walmart if looking for ways to accurately classify the trips using the purchase data of a customer.

By providing Walmart an approach to accurately classify the trip based on customer's purchase data, Walmart will be able to calculate the trip types on the go instead of relying on off line and delayed batch processing resulting in a quicker and near instantaneous analysis of the customers' buying patterns leading to

- Happy customers as there will be less chance that they don't find what they are looking for
- Easier store management and operation by providing quicker insights to the store manager on how to stock the store's inventory
- Improved revenue by analyzing trends faster and reacting to change in customer behavior

1.3. APPROACH

The high-level steps to solve this problem will be

1. Data cleansing and wrangling
2. Deciding metrics to be used to calculate the accuracy
3. Applying various Machine Learning (ML) algorithms to the cleansed data to find the optimal solution

The overall solution would be an iterative process, with multiple cycles of data cleansing and application of algorithms happening throughout the process until an optimal solution with satisfactory accuracy is arrived at.

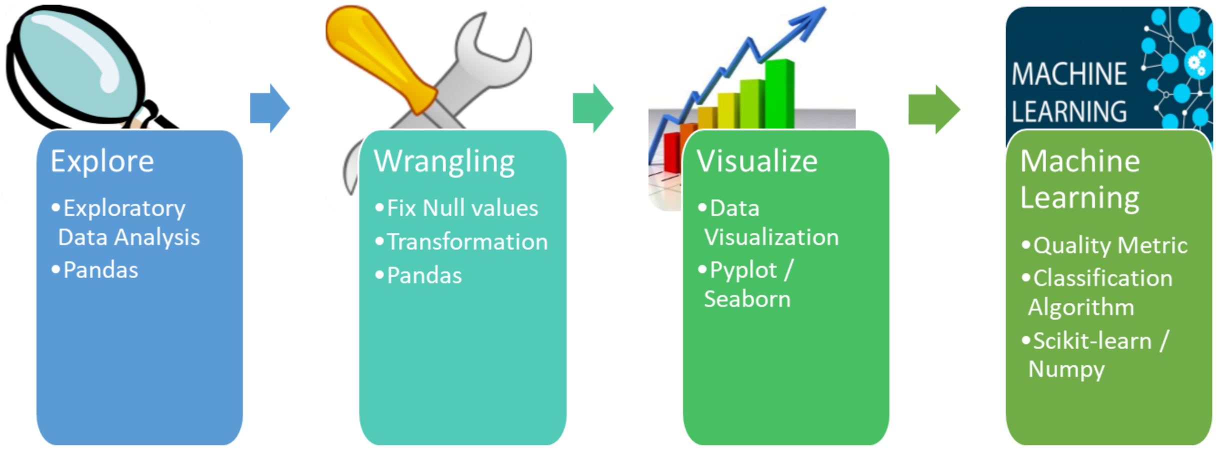


FIGURE 1 SOLUTION APPROACH

1.4. DELIVERABLES

Deliverables of this project will include

1. Code – iPython notebook. Code with relevant documentation included.
2. Report – a MS Word document that describes in detail the approach, the findings, the final result and recommendations
3. Presentation – a MS PowerPoint slides that presents the problem and results

2. About the Data

2.1. Data Used

The data to be used for this will be from the competition page. The site provides training and testing data. Training data consists of 647054 observations of purchase. The purchase data contains

- TripType - a categorical id representing the type of shopping trip the customer made. This is the ground truth that you are predicting. TripType_999 is an "other" category.
- VisitNumber - an id corresponding to a single trip by a single customer
- Weekday - the weekday of the trip
- UPC - the Universal Product Code (UPC) number of the product that is being bought or returned
- ScanCount - the number of the given item that was purchased. A negative value indicates a product return.
- DepartmentDescription - a high-level description of the item's department
- FinelineNumber - a more refined category for each of the products, created by Walmart

2.2. Dataset Characteristics

The data that has been provided has a large number of order details, has the right data is the right format in most of the observations and has less need to be cleaned. There is a total of approx. 647 K invoice items in this dataset. Below are the important dimensions of this dataset.

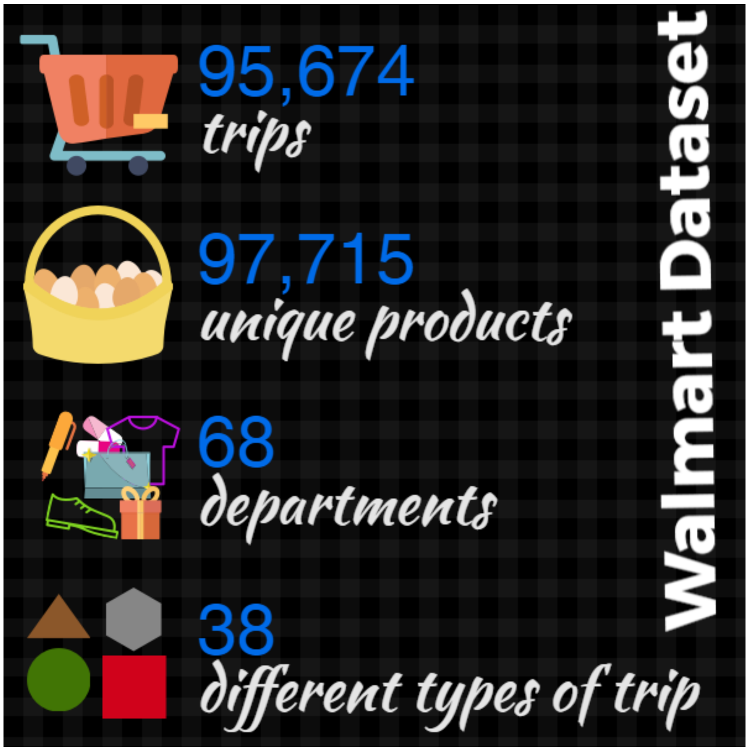


FIGURE 2 WALMART DATASET CHARACTERISTICS

Highlights of the nature of this data

- The number of unique products (represented by the UPC) is large compared to the number of the trips in the dataset
- Roughly 10% of the UPCs account for more than 50% of the transactions
- The same behavior can be seen with the Department as well

3. Solution Details

3.1. Data Cleaning / Wrangling

The given dataset is fairly clean with not much of missing values / information. The only data cleanup that had to be performed was the check for presence of null values in the UPC or the Department, since these are incomplete observations and cannot be used to classify the Trip Type. In total, there were around 1,500 (0.2 %) observations with null UPC / Department. There are some observations with department that has a valid value ("PHARMACY") with varying trip types. For this exercise, the assumption is that these records will null UPCs or Departments have been created in error and will be dropped.

The objective of this model is to classify the shopping trip. Many individual transactions (buy or returns) make a shopping trip, the current data set consists of individual observations where each product bought or returned is tied to a visit to the store. In order to classify the type of the visit, this provided data needs to be aggregated to have all the details (products purchased / returned, quantity, etc.) of a trip as one single observation. The below represents this transformation.

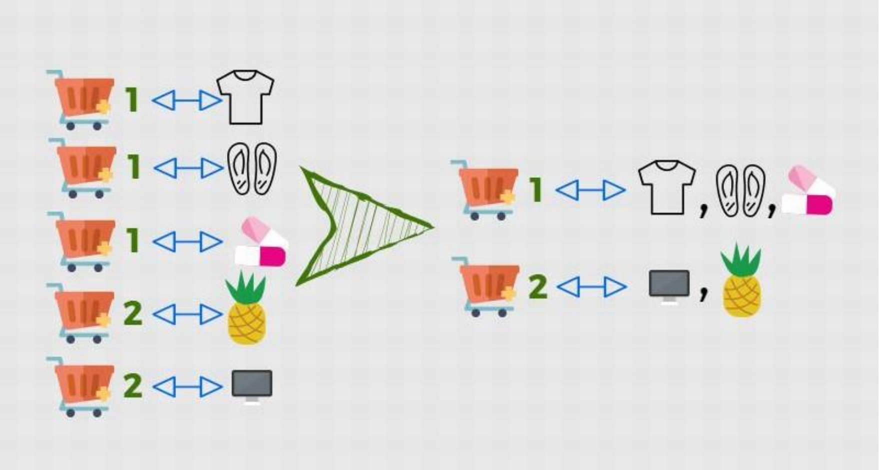
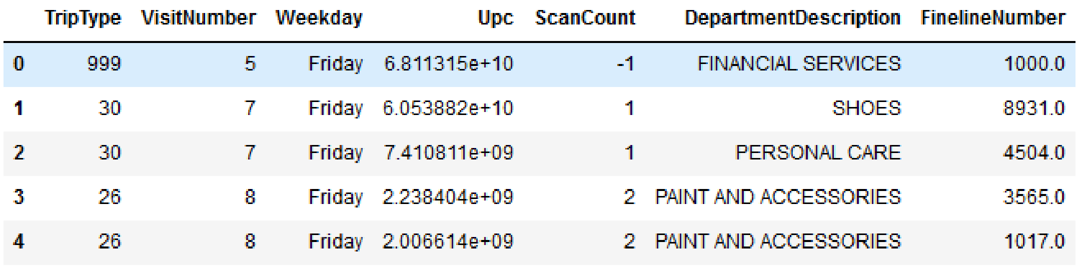


FIGURE 3 DATA WRANGLING APPROACH

To make this aggregation, I have used `Pandas.DataFrame.groupby` function. The dataset after removing the null UPC / Department values was grouped by the VisitNumber to collect information for each visit and store it along with the visit in a new DataFrame.

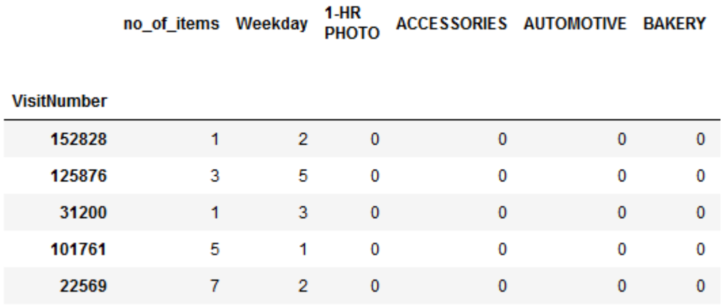
Below is a snapshot of original data provided



	TripType	VisitNumber	Weekday	Upc	ScanCount	DepartmentDescription	FinelineNumber
0	999	5	Friday	6.811315e+10	-1	FINANCIAL SERVICES	1000.0
1	30	7	Friday	6.053882e+10	1	SHOES	8931.0
2	30	7	Friday	7.410811e+09	1	PERSONAL CARE	4504.0
3	26	8	Friday	2.238404e+09	2	PAINT AND ACCESSORIES	3565.0
4	26	8	Friday	2.006614e+09	2	PAINT AND ACCESSORIES	1017.0

FIGURE 4 ORIGINAL DATA

Below is a snapshot of the data that will be input to the Machine Learning algorithm



	no_of_items	Weekday	1-HR PHOTO	ACCESSORIES	AUTOMOTIVE	BAKERY
VisitNumber						
152828	1	2	0	0	0	0
125876	3	5	0	0	0	0
31200	1	3	0	0	0	0
101761	5	1	0	0	0	0
22569	7	2	0	0	0	0

FIGURE 5 TRANSFORMED DATA

3.2. Exploratory Data Analysis

Inferential Statistics

Are there variables that are particularly significant in terms of explaining the answer to your project question?

Section 1.1. gives details on the variables / columns contained in the provided dataset. The important variables that are

1. The department of the product
2. UPC code of the product
3. Fineline number of the product
4. Total number of products bought / scanned
5. Day of the week

The objective of this exercise / problem is to create an algorithm that can predict the trip type given the above information. Intuitively, the trip type will depend on the product being transacted (bought or returned) during the shopping trip. However, the number of products is huge (around 97,000), so, the next best indicator would be the department to which the product belongs.

Are there strong correlations between pairs of independent variables, or between an independent and a dependent variable?

As part of the data wrangling to setup the data, the total number of items bought / scanned in a shopping trip was calculated for each unique visit. The below chart shows the correlation between the Trip Type and the total number of items scanned

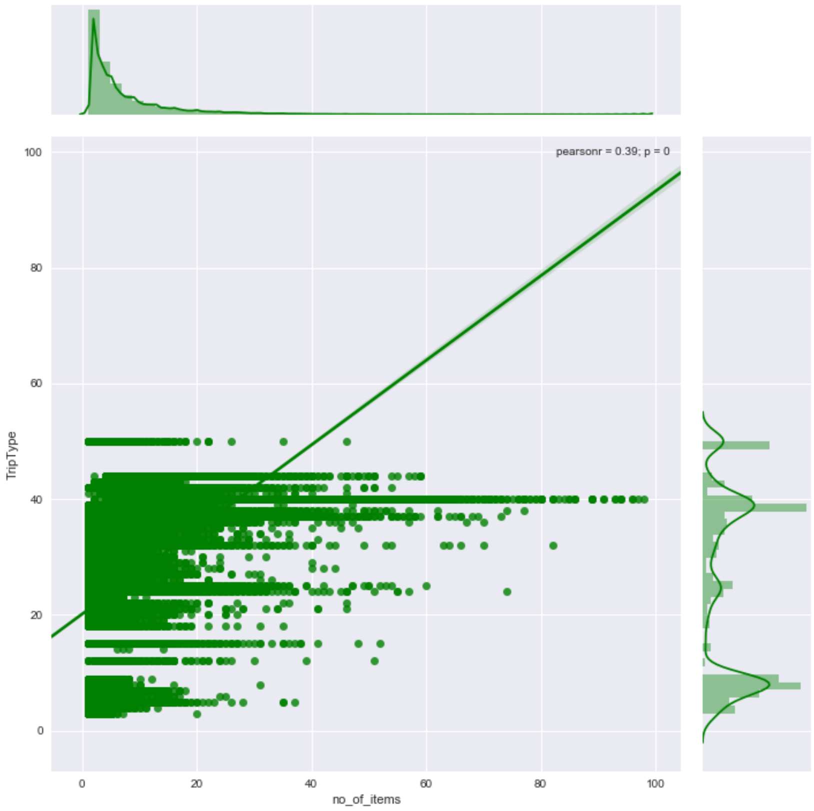


FIGURE 6 TRIP TYPE - NUMBER OF ITEMS CORRELATION

From the chart it visually appears that there may be no clear / meaningful correlation between the Trip Types and the number of items scanned. There are some trip types (trip types that are '20' and greater) that only when the number of items scanned is > 40. This will have to be confirmed by the classification algorithms used as part of the solution.

The next correlation that can be checked is if the Trip Types are related to the day of the week. The below charts show the distribution of the trip types for each day of the week.

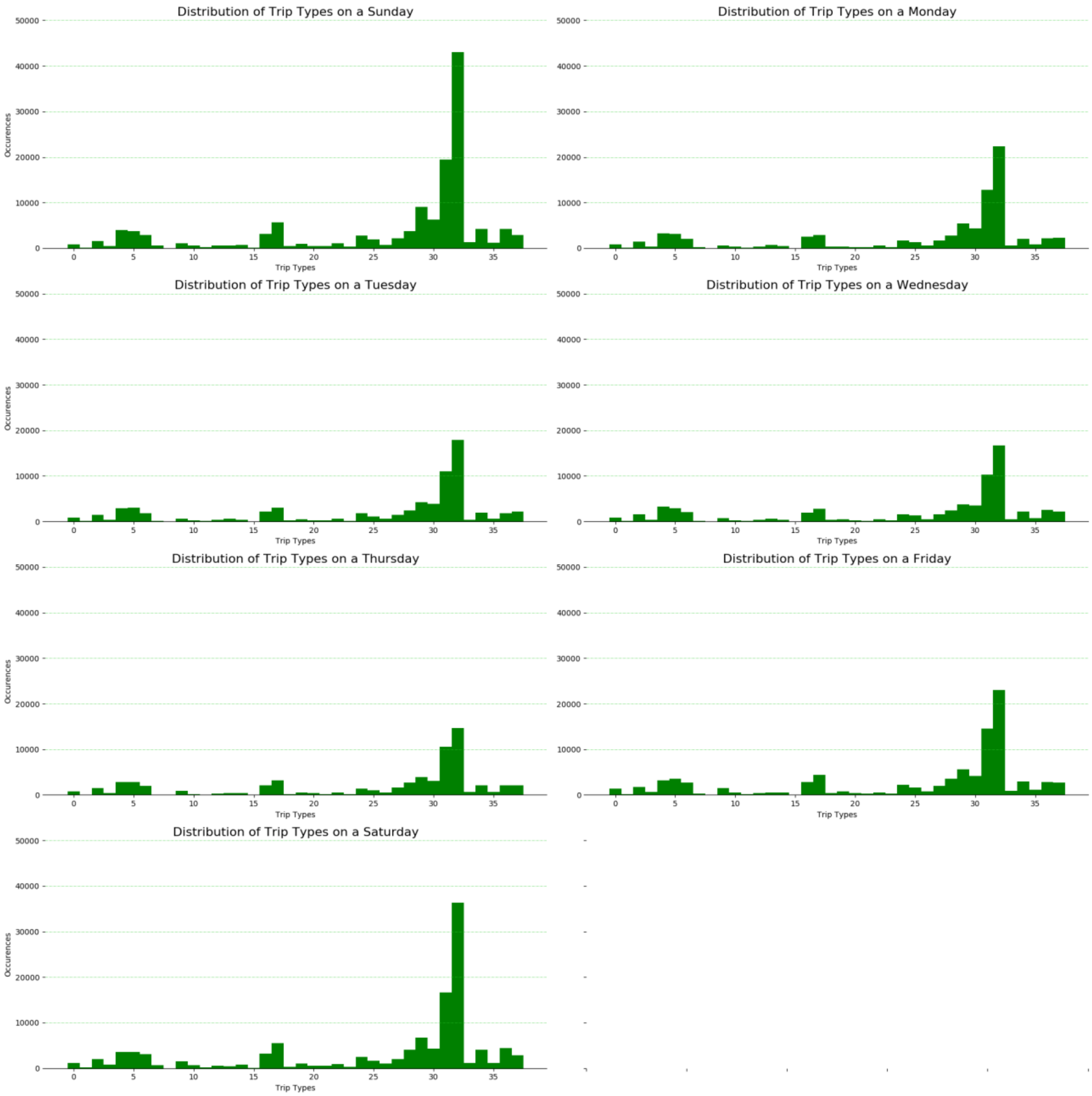


FIGURE 7 TRIP TYPES ACROSS DAYS OF A WEEK

From the above charts, one can see that a similar pattern of distribution of the Trip Types are seen on all days of the week. These charts also show that the traffic to the store is high on Weekends and is low during the Weekdays, except on Fridays. Sunday is the most popular day and, Thursday the least popular.

3.3. Data Visualization

As explained in section 2.1. the distribution of Departments and UPCs are not uniform, i.e. relatively small number of Departments occur in a large number of observations. For example, items from Grocery Dry Goods have been scanned in over 70,000 transactions, that represents more than 7% of the total transactions. The below chart shows the total items scanned by each department. As you can see in the bottom half the chart, there are more than half the departments that have just a few thousand transactions.

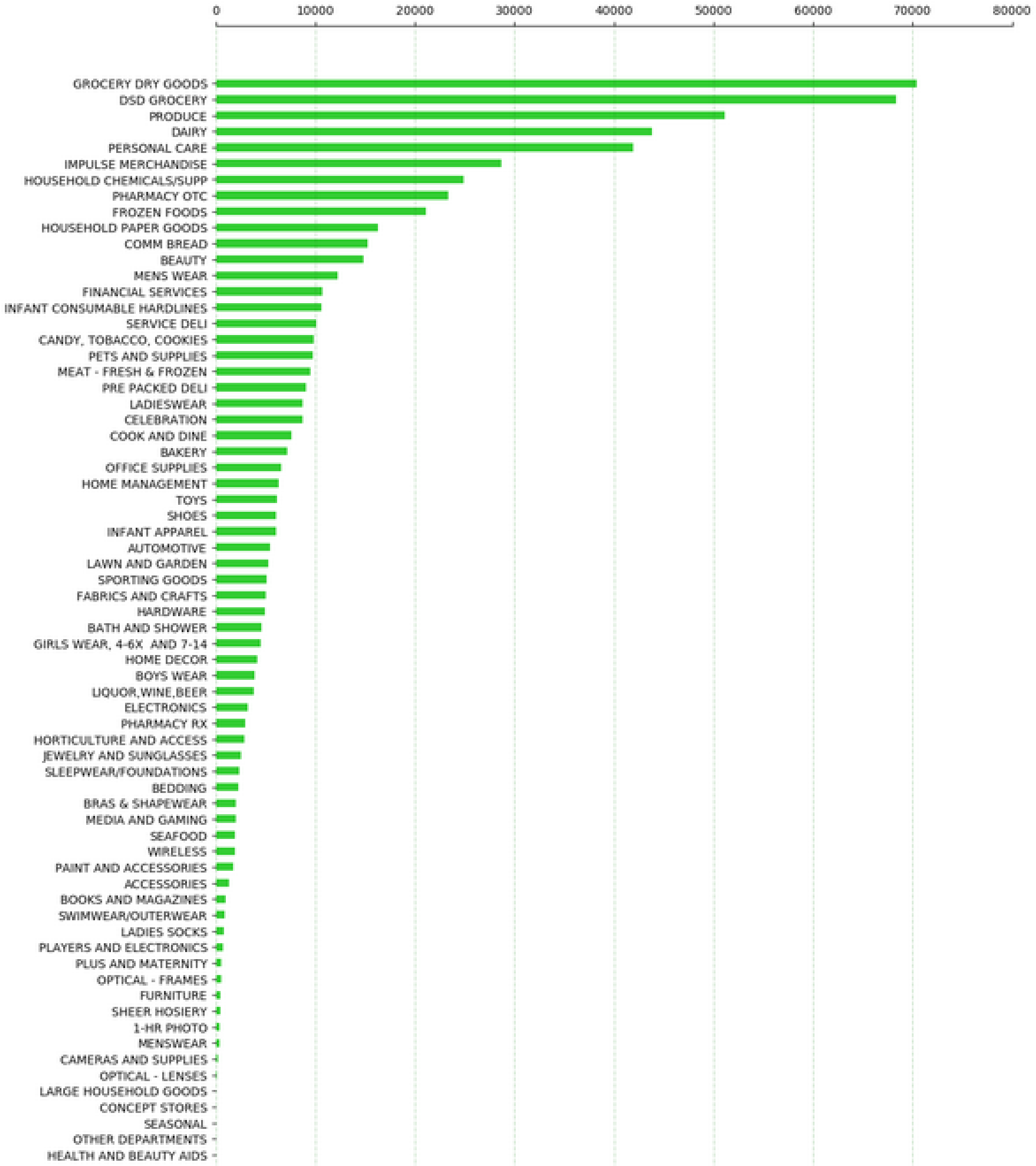


FIGURE 8 TOTAL ITEMS SCANNED BY EACH DEPARTMENT

Below is another visualization of the same information. This shows that the top 20% of the Departments account for close to 80% of the transactions.

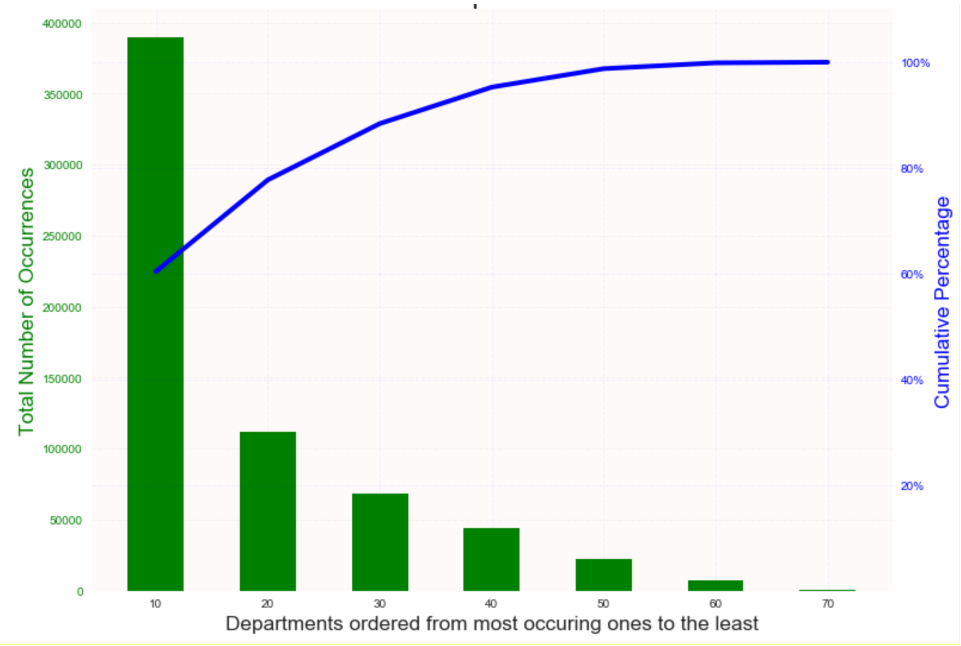


FIGURE 9 DISTRIBUTION OF DEPARTMENTS IN THE DATA

Similar behavior is seen in the distribution of the UPC numbers as seen in the below chart. 10% (10,000) of the UPCs account for close to 60% (approx. 390,000) of the observations.

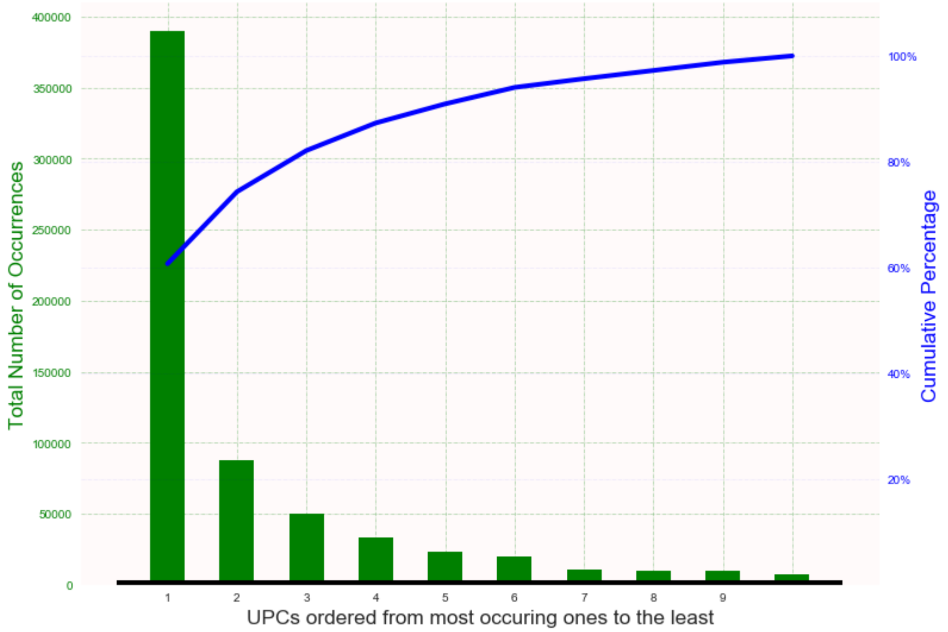


FIGURE 10 DISTRIBUTION OF UPC IN THE DATA

3.4. FEATURE SELECTION

To classify the type of the visit, following are the potential features that can be used

1. Total number of items scanned

2. Day of the Week
3. Product Department
4. Product code – UPC
5. Fineline Number

The provided data is in the form of individual item scans. So, we will have to associate all the items bought in one trip to that particular trip.

The data size of number of items scanned and the day of the Week are small to be used as features as-is, however, UPC number and Fineline Number are categorical data and have a large number of categories. There are 97,715 UPC numbers in total. It will not be feasible to run machine learning (ML) algorithms to run a dataset with 100 K features. So, I will be using the top 5% of the UPC numbers. Below are the features that I used in my models

1. Total number of items scanned
2. Weekday of the trip
3. Departments
4. Top 5% of the UPC numbers
5. Top 3% of the Fineline numbers

The total number of features for the model is 6514.

3.5. Quality Metric

I plan to use accuracy as a quality metric to determine the model performance. The accuracy calculated will be as a % of accurate prediction of the test set. This is simple, straightforward and will be given by

$$\frac{Total\ Number\ of\ Correct\ Predictions}{Total\ Predictions}$$

3.6. Machine Learning

This is a classification problem, below are the ML algorithms that I will be running. First, I will run these algorithms with default values, then tune the hyper-parameters of the model to achieve the best model. The total number of observations (Visits) used were 94247, each had 6514 features. The train-test split was taken to be 80/20.

Below are the models that I have tried

1. Logistic Regression - In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function.
2. Stochastic Gradient Descent (SGD) Classifier - a plain stochastic gradient descent learning routine which supports different loss functions and penalties for classification.
3. SVC – Support Vector Classification
4. LinearSVC - Linear Support Vector Classification
5. Gaussian Naïve Bayes – implements the Gaussian Naive Bayes algorithm for classification
6. Decision Trees - non-parametric supervised learning method used for classification and regression.
7. Random Forest Trees – In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set.

The models had varied performance and accuracy. Below table lists the time taken by and the accuracy of all models.

Algorithm	Time Taken to fit the model	Accuracy on Test dataset
Logistic Regression	4h 35min 50s	66.32 %
SGD Classifier	2min 32s	61.24 %
SVC	11h 36min 41s	43.94 %
Linear SVC	2min 18s	62.12 %
Gaussian NB	16.4 s	20.69 %
Decision Trees	1min 4s	61.75 %
Random Forests	1min 36s	67.75 %

FIGURE 11 ML MODEL PERFORMANCE COMPARISON

The Random Forests algorithm had the best accuracy, 67.75 %, and it took a little more than 1.5 minutes to fit the data. This algorithm was further analyzed to find the best hyper-parameters. The parameters that were analyzed were

- `n_estimators` - The number of trees in the forest. The best value was found to be 60.
- `max_features` - The number of features to consider when looking for the best split. The best value was "auto", i.e. max_features=sqrt(n_features).
- `max_depth` - The maximum depth of the tree. The best value was 2200
- `n_jobs` - The number of jobs to run in parallel for both fit and predict. This was set to -1, this sets the number of jobs to the number of cores.

By tuning the model based on the above parameters, the *accuracy on test data increased to 68.18%.*

The most important factors that determine the classification

Based on the model, the most important features that determines the Trip Type are

1. no_of_items
2. DSD GROCERY
3. Weekday
4. PHARMACY OTC
5. PERSONAL CARE
6. GROCERY DRY GOODS
7. MENS WEAR
8. DAIRY
9. PRODUCE
10. SERVICE DELI

This is in line with intuition that the major factor that would determine the Trip is the Department and not the UPC number of the product. The departments that appear most are given higher weightage in the trip type decision.

The accuracy of the model can be improved further. The below confusion matrix shows the areas where this model gets the results incorrect

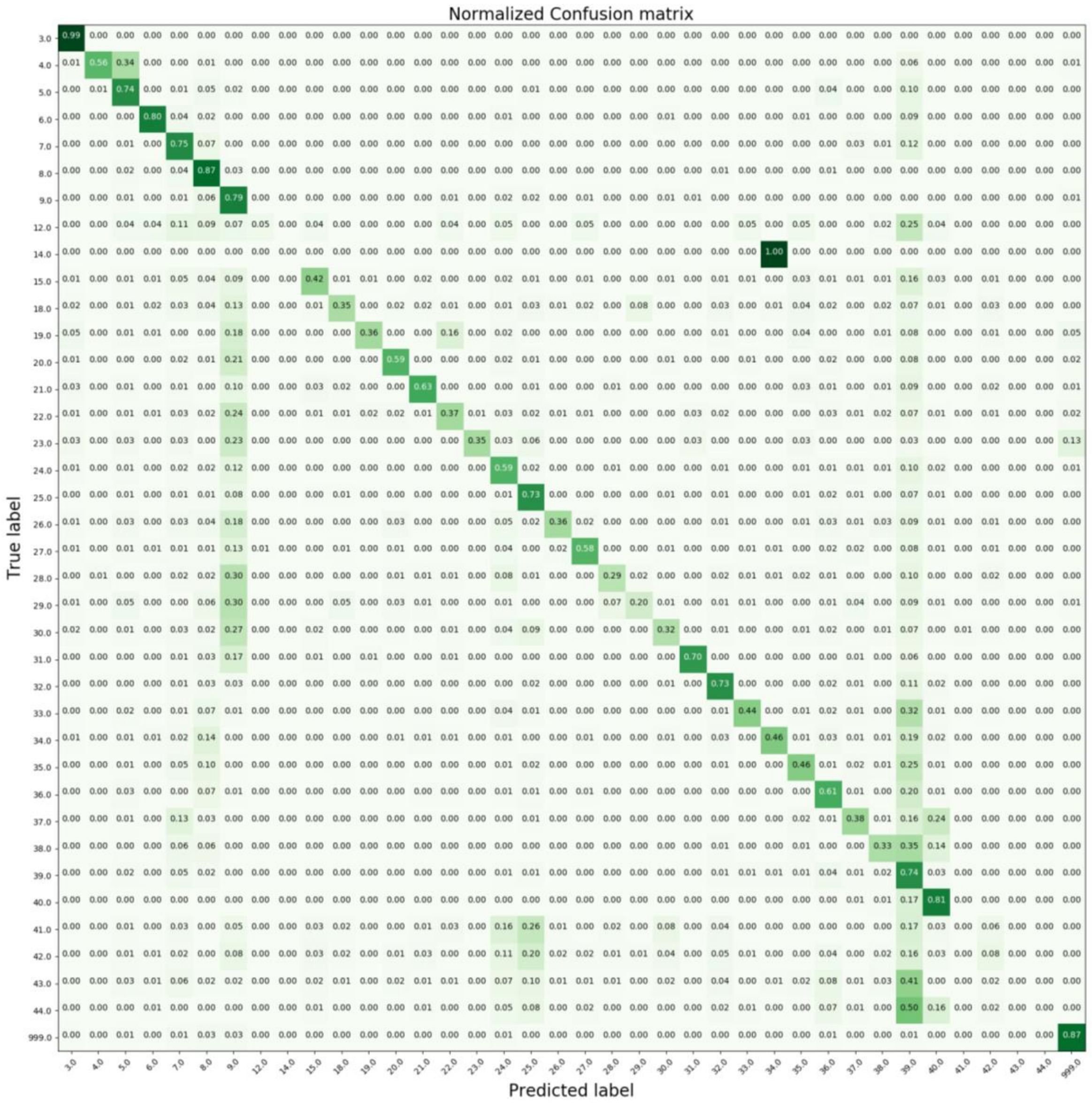


FIGURE 12 NORMALIZED CONFUSION MATRIX

For example, from the confusion matrix, we can see that this model has very bad prediction for Trip Types 12 and 14 . This model performs well for the first few Trip Types. Further analysis needs to be done to tune the model.

4. Source Code

4.1. Git Repository

The source code for this project is maintained on Git at: https://github.com/rajeshdsar/Walmart-Kaggle-Comp-Repo