Springboard Data Science Career Track Capstone Project I

Walmart Trip Type Classification

By Rajesh Dharmarajan

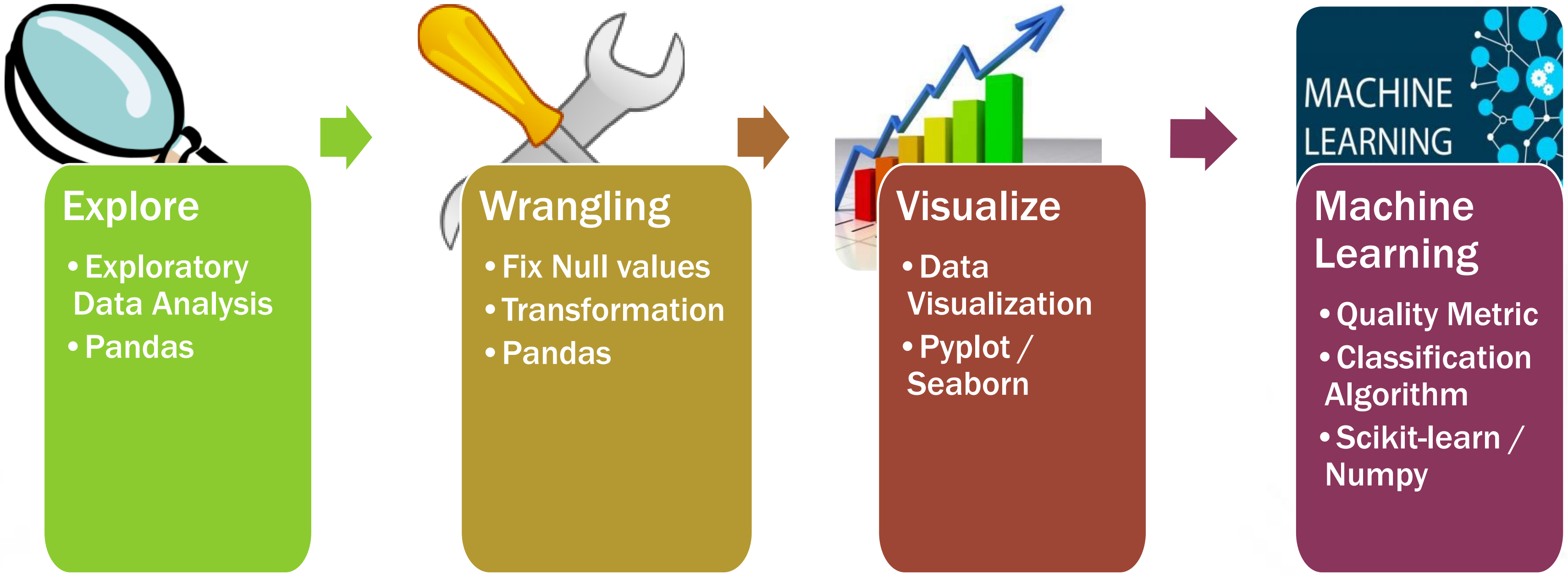
Table of Contents

- Problem Overview
- Summary of Approach and Results
- Solution Details
  - Data Wrangling
  - Data Visualization
  - Machine Learning
- Summary
- Future Implementation Suggestions

Walmart Shopping Trip Type Classification

- Walmart improves customers' shopping experiences by **segmenting their store visits into different trip types**
- Challenge : Accurately classify customer trips using only a transactional dataset of the items they've purchased
- The transactional data provided contains information about – the items bought / returned, department to which the item belongs and unique codes that identify the item
- Large number of transactions provided

Summary of Approach



Explore

- Exploratory Data Analysis
- Pandas

Wrangling

- Fix Null values
- Transformation
- Pandas

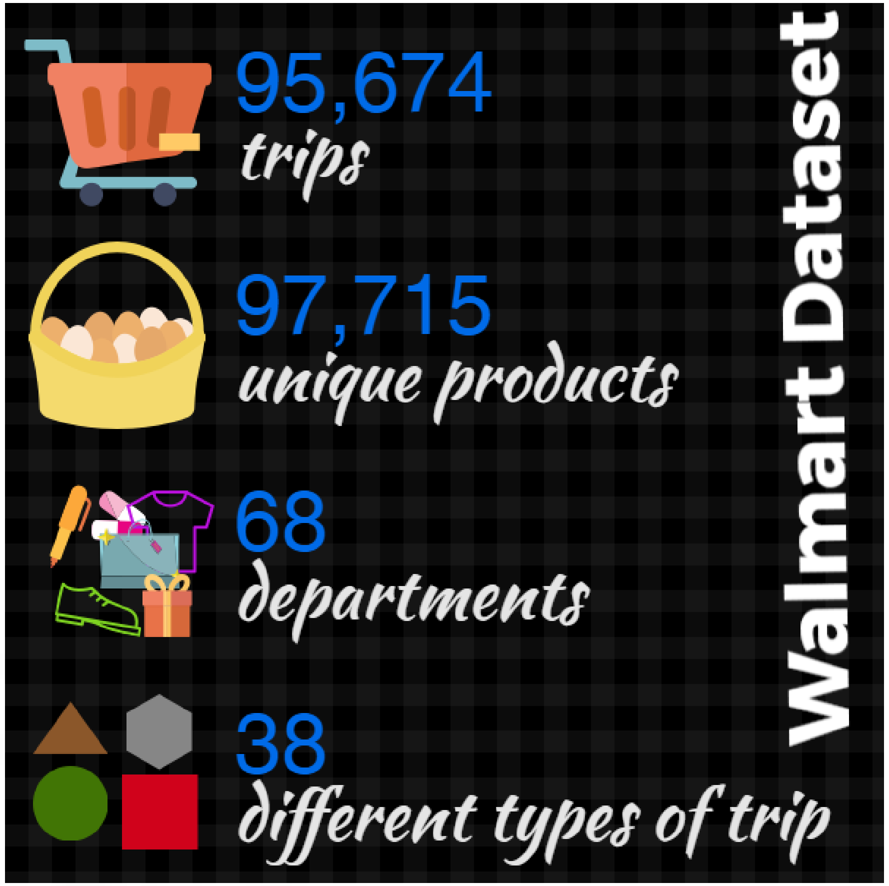
Visualize

- Data Visualization
- Pyplot / Seaborn

Machine Learning

- Quality Metric
- Classification Algorithm
- Scikit-learn / Numpy

Dataset Characteristic



Walmart Dataset

95,674 *trips*

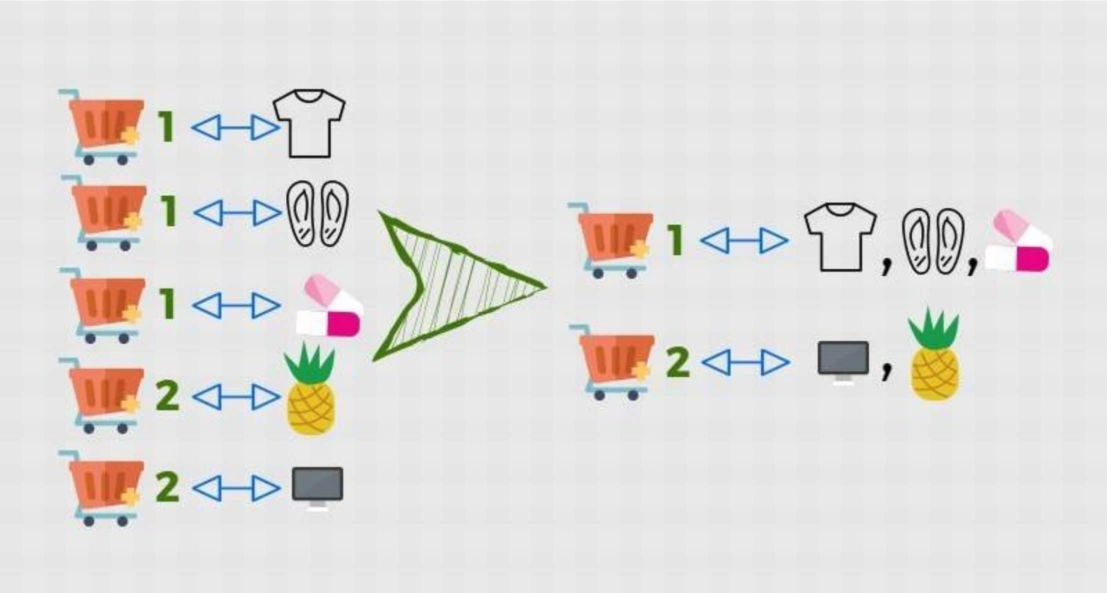
97,715 *unique products*

68 *departments*

38 *different types of trip*

- Large number of data provided
- Provided data is clean, very few observations (0.2 %) have null UPC numbers or Departments
- The records with null were dropped
- Relatively lower number (20 to 25%) of Departments and UPCs appear in large number (approx. 80%) of

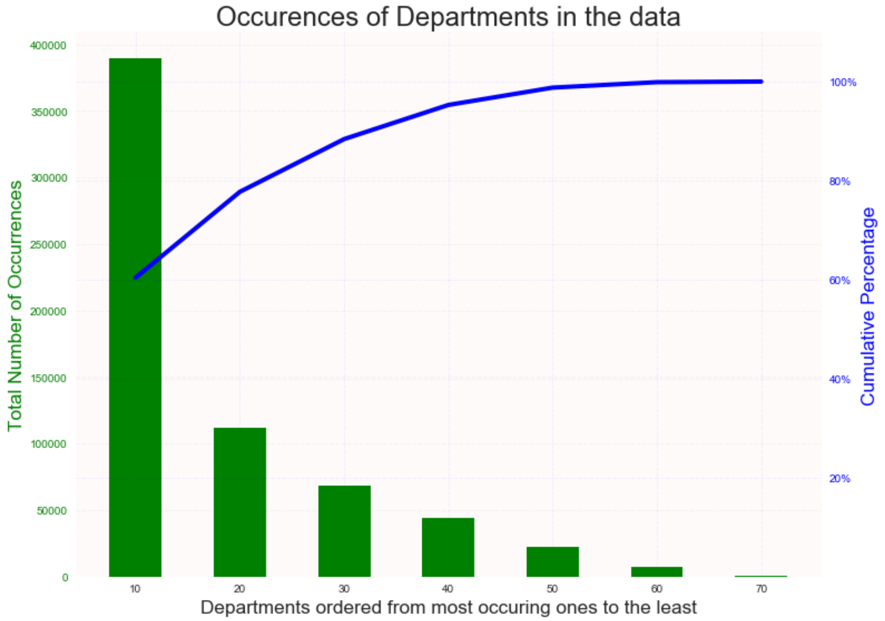
Data Wrangling



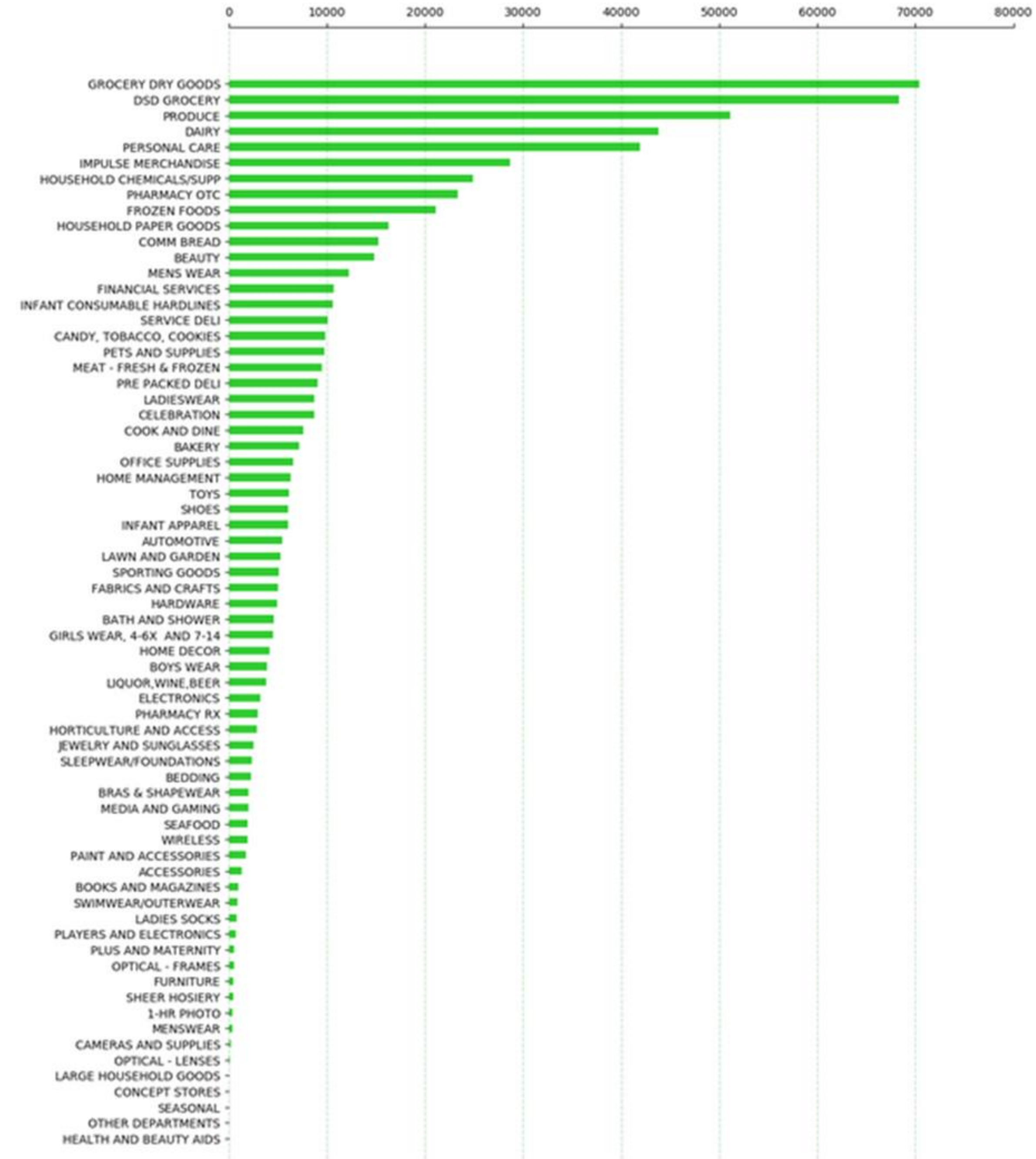
- the provided data set consists of individual observations where each product bought or returned is tied to a visit to the store

- This provided data needs to be aggregated to have all the details (products purchased / returned, quantity, etc.) of a trip as one single observation

Data Visualization



- Relatively lower number (20 to 25%) of Departments and UPCs appear in large number (approx. 80%) of observations
- Similar behavior seen in UPC numbers and Fineline numbers



Data Visualization

- Grocery, Produce, Dairy are the most bought products
- Each of these occur in over 50,000 transactions
- Seasonal and Health / Beauty do not figure in the list, possibly due to the time period when this data was collected

Machine Learning Models Used

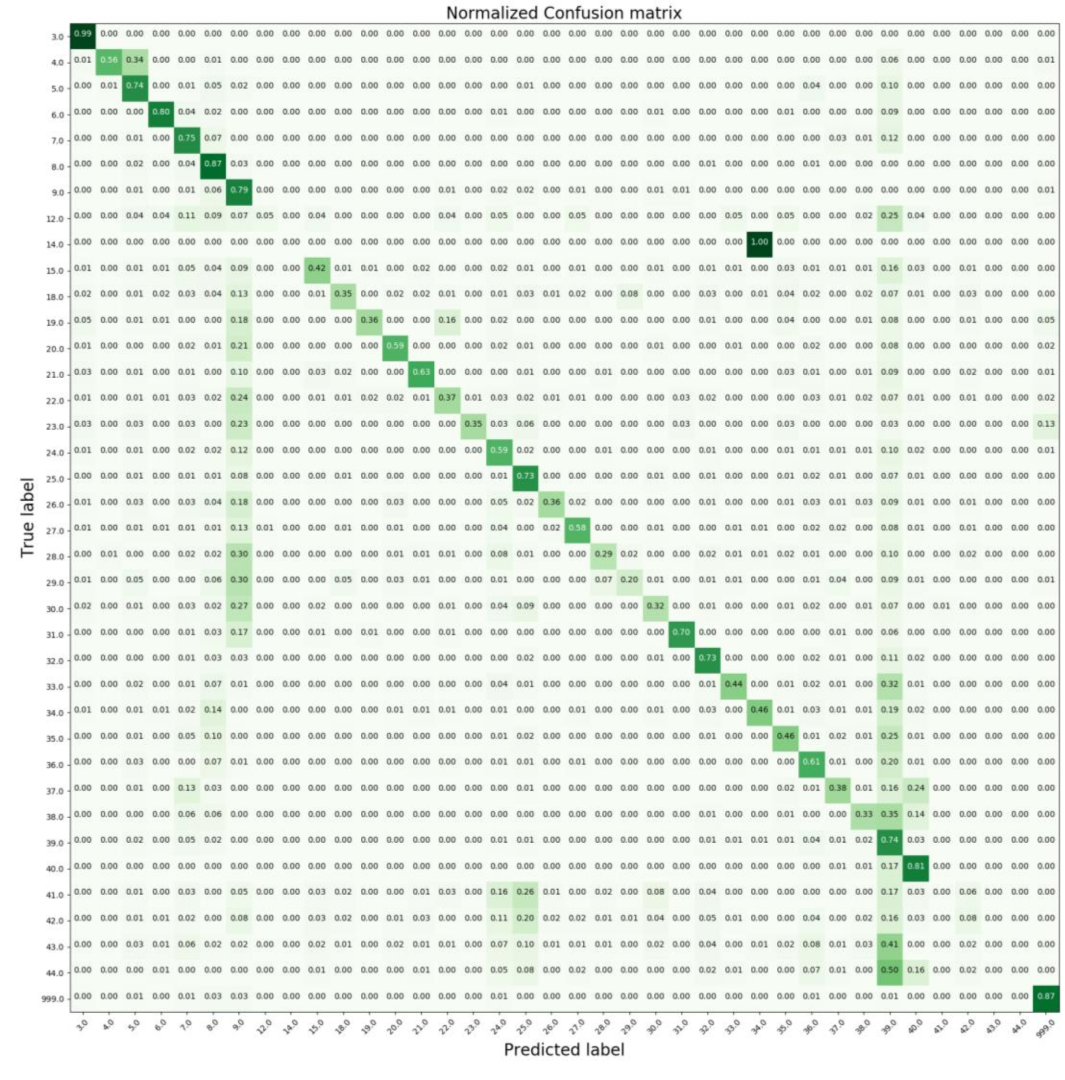
Algorithm	Time Taken to fit the model	Accuracy on Test dataset
Logistic Regression	4h 35min 50s	66.32 %
SGD Classifier	2min 32s	61.24 %
SVC	11h 36min 41s	43.94 %
Linear SVC	2min 18s	62.12 %
Gaussian NB	**16.4 s**	20.69 %
Decision Trees	1min 4s	61.75 %
Random Forests	1min 36s	**67.75 %**

- Train-test data taken at 80/20
- Accuracy of prediction used as the metric to evaluate the models
- The Gaussian Naïve Bayes model took the least amount of time, however, it had very poor accuracy
- Random Forests produced the best accuracy

The Most Important Features

- 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

Confusion Matrix



- The accuracy of the model can be improved further
- 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