Capstone Project: Predict Walmart Trip Types

The problem:

Walmart is trying to segment different types of customers based on different 'trip types' their customers have taken. We can utilize a deep learning algorithm to understand which factors 'predict' whether a customer will be classified in particular 'trip type'.

Data Cleaning and Feature Engineering

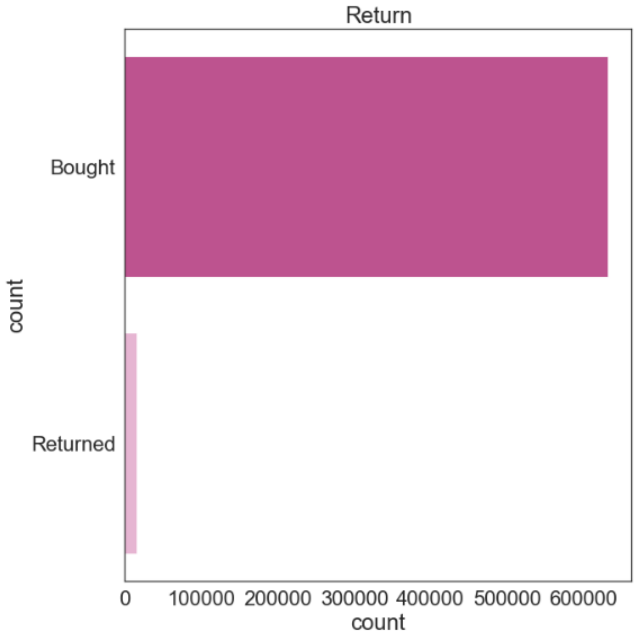
General Data Analysis:
Several data analytics steps were taken to preprocess data. First, numerous new features were added such as: Return, Count_of_Item, Rename_Dep, and Commodity. The engineered variables and their description were added in the table below.

Variable	Description	Type
Visit Number	When they visited	Original Data
---	---	---
Weekday	Day when the visit occured	Original Data
---	---	---
Upc	Code printed on retail products	Original Data
---	---	---
ScanCount	Items Scanned or Returned	Original Data
---	---	---
DepartmentDescription	Name of department (granular)	Original Data
---	---	---
Rename_Dep	Name of department (aggregated)	Engineered
---	---	---
Return	If items was returned or bought	Engineered
---	---	---
Count_of_item	Number of items bought - single or multiple	Engineered
---	---	---
COMMODITY	PLU item name	Engineered

Overall Exploratory Data Analysis:

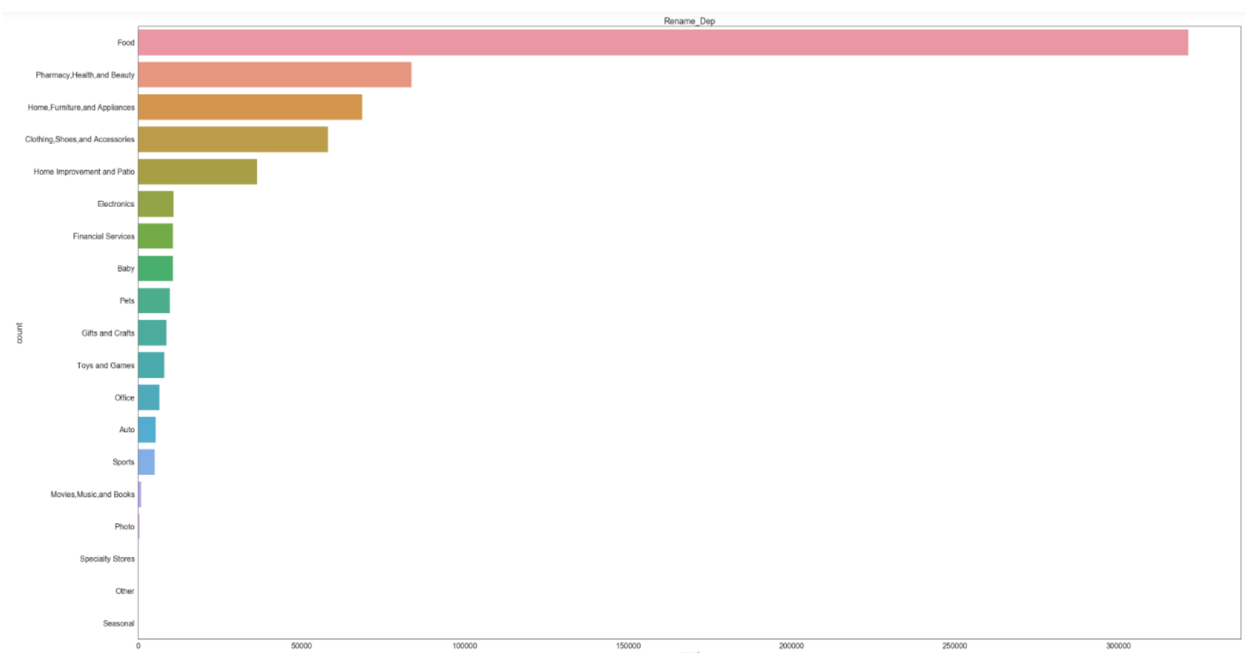
1.1 Returned Item

Based on analysis, it seems like most of the items were bought with very few items being returned.



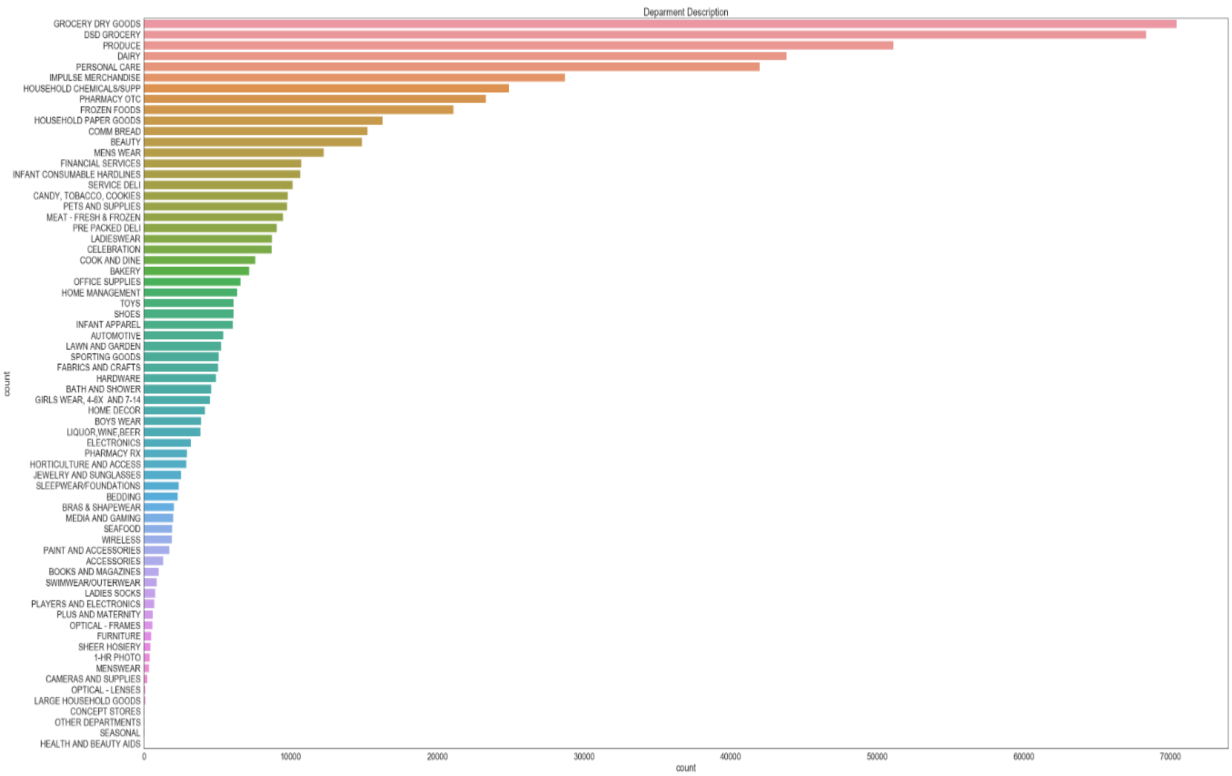
1.2 Department Description (Aggregated)

It seems like most of the items that were bought are from the Food Department with Pharmacy, Health, and Beauty being second.



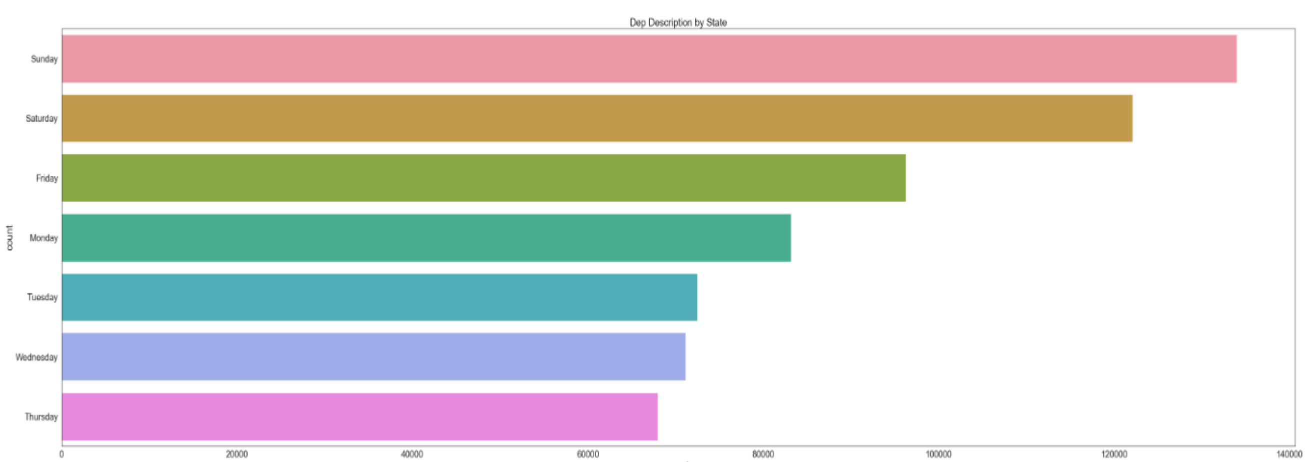
1.3 Department Description (Granular)

Instead, if we look at the departments on a granular level, we could see that most bought items come from the Department of Grocery Dry Goods, Produce, and Diary indicating that consumers are more likely to purchase items from those departments.



1.4 Weekday Analysis

In contrast, most of the items are bought during the weekend (Sunday and Saturday).



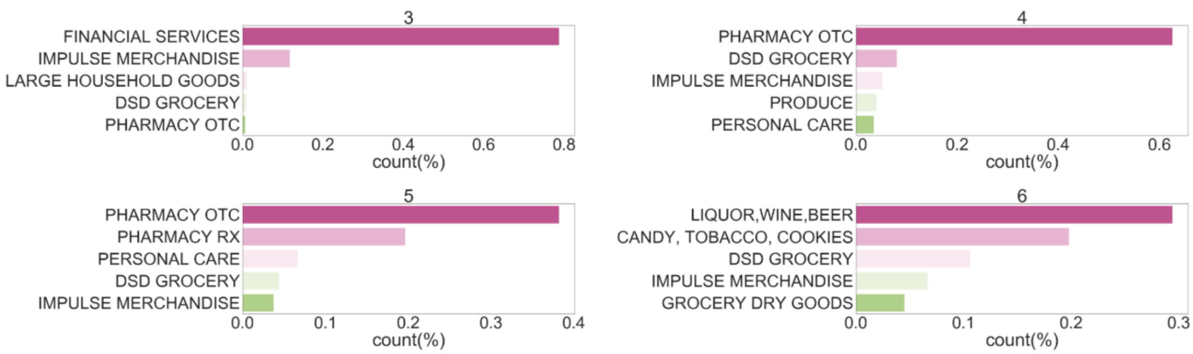
Summary:

Overall, this analysis paints the following picture: majority of the items are bought during the weekend and are generally from the Grocery and Produce departments. Thus, this indicates that most of the audience are people who perhaps go to Walmart for their weekly Grocery shopping.

Exploratory Analysis by Trip Type

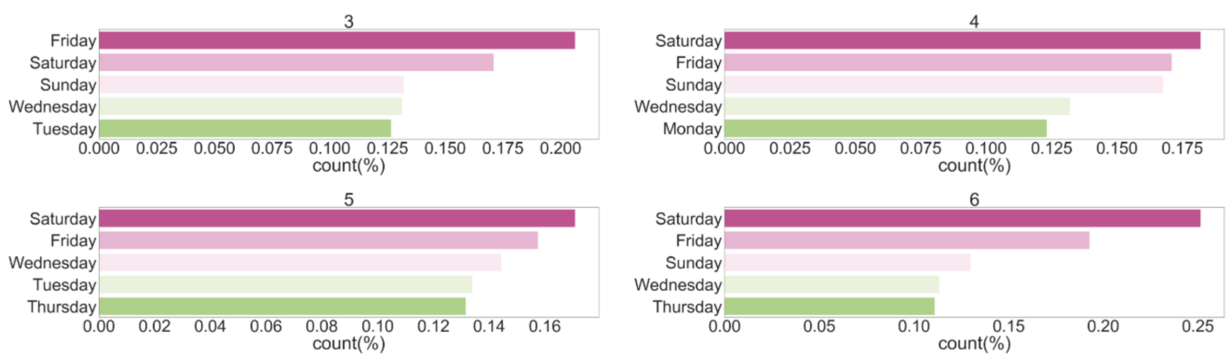
1.1 Department Description by Trip Type

Per analysis below, it seems like TripType varies by Department Description. For example, for Trip 3-6, you would see that Trip 6 is more composed of people who buy alcohol and candy (maybe, a special occasion due to Valentine's day) while Trip 4 is more related to Pharmacy and DSD grocery.



1.2 Weekday by Trip Type

The analysis below shows some differences in weekdays in different trip types. In the extract below, Trip Type 3 most of the items were bought on Friday while for Trip Type 4-6 most of the items were bought on Saturday.



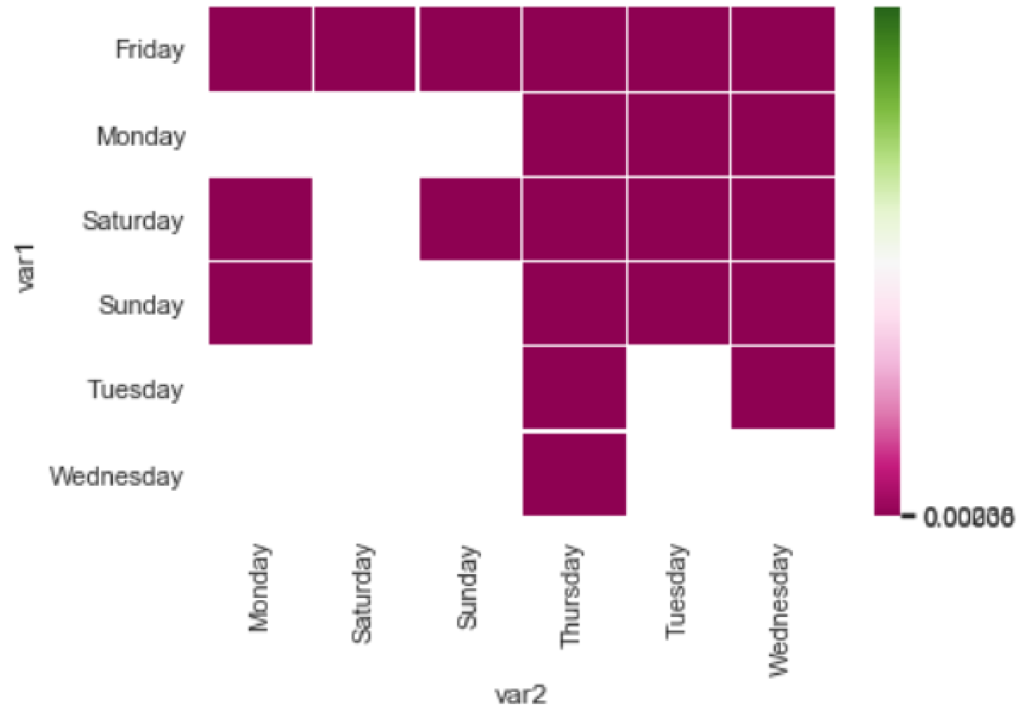
Statistical Analysis

Given that weekdays and departments vary by Trip Type, we might investigate whether there are statistically significant differences. Using post-hoc analysis with Bonferroni correction, we investigated whether there are statistical differences between different departments and weekdays.

1.1 Department

It seems the number of products bought from Financial Services is significantly different compared to the number of products bought from most other departments with exception of products from Baby and Electronics. Similarly, number of items bought Specialty Stores department is significantly different than most other departments with exception of Seasonal and Other.

1.2 Weekdays



It seems like the number of items bought or returned are significantly different from other days of the week.

Machine Learning and Engineering

In-depth Model Selection

For this model, the objective is to predict multiple Trip Types which is a multiple classification problem. Given that the dataset is large and there are many variables with high cardinality, a deep learning neural network was attempted. Thus, two separate neural networks were attempted. The first one was a simple neural network that aggregated data across Visit Number. The second model treated each Visit Number as a sequence of values.

Model Architecture

This project tried to predict the trip type using two separate models. First of them, was a neural network with a single layer (1000 neurons) while the second was a recurrent neural network with a single layer (600 layers). In addition to that, both models included L1 regularization and early stopping.

Model Data Cleaning

As we are dealing with two separate models, there were several preprocessing steps that were taken.

For the simple neural network, each variable in the model was one hot encoded. For variables with high sparsity (Upc, Fineline Number) only the top 50% values with highest frequency were kept. Furthermore, after all variables were one-hot encoded and a single dataset was created, the dataset was aggregated across Visit Number. So, the one hot encoded variable were summed across each visit number.

For the recurrent neural network, the data was again aggregated by Visit Number; however, instead of the data being summed, each row contained the first 15 values for each column. For example, for each Visit Number, one row could contain that first 15 departments they visited during their trip. As we have multiple columns, each timestamp contains multiple features.

Model Results

The first model of single-layer network had an accuracy of 79% while the recurrent neural network had an accuracy of 22%. Thus, this analysis will only consider the simple single-layer network moving forward.

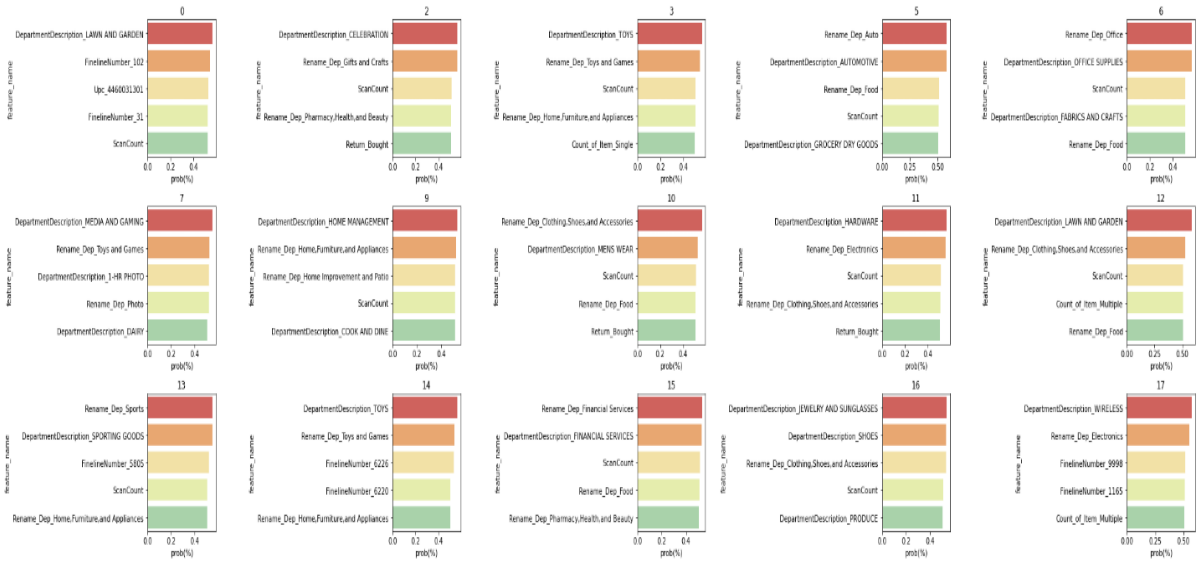
Precision and Recall

Although Precision and recall varies by Trip Type, there are some Trip Types which precision and recall are significantly worse than chance. For example, Trip Type 28 to 31 have lower precision and recall, maybe indicating that the model cannot distinguish between Trip Types.

	precision	recall	f1-score	support
0	0.52	0.36	0.42	45
2	0.61	0.50	0.55	135
3	0.65	0.60	0.62	58
4	0.50	0.39	0.44	18
5	0.62	0.75	0.68	60
6	0.76	0.68	0.72	69
7	0.54	0.58	0.56	60
8	0.00	0.00	0.00	5
9	0.71	0.76	0.73	282
10	0.77	0.79	0.78	375
11	0.68	0.63	0.65	51
12	0.70	0.73	0.71	73
13	0.65	0.68	0.67	50
14	0.64	0.42	0.51	43
15	0.98	0.97	0.97	722
16	0.52	0.55	0.53	86
17	0.78	0.85	0.82	73
18	0.75	0.79	0.77	254
19	0.77	0.79	0.78	219
20	0.65	0.85	0.74	65
21	0.77	0.72	0.75	387
22	0.78	0.82	0.80	357
23	0.71	0.69	0.70	531
24	0.76	0.78	0.77	546
25	0.00	0.00	0.00	1
26	0.78	0.47	0.58	30
27	0.85	0.87	0.86	1192
28	0.28	0.14	0.18	80
29	0.46	0.43	0.45	240
30	0.36	0.31	0.33	135
31	0.61	0.47	0.53	242
32	0.85	0.91	0.88	514
33	0.75	0.82	0.78	109
34	0.81	0.87	0.84	915
35	0.91	0.89	0.90	1376
36	0.73	0.82	0.77	353
37	0.93	0.89	0.91	613
accuracy			0.79	10364
macro avg	0.65	0.64	0.64	10364
weighted avg	0.79	0.79	0.79	10364

Model Interpretability

Using SHAP values, we investigated determining which variables explain the model results.



In the table above that showcases the top 5 most important variables for the first 13 variables. As you see, for Trip Type 5 the auto and food department were important factors. While for Trip Type 3, single item purchases from the toys department are important predictors.

Next Steps

Use information from the model interpretability to test results from neural nets
For each trip type, look at the top features and create tests:

- Walmart can a) put the two departments close by or b) create offers where you get each an item from each department for the price of 1
- See if those type of discounts result in higher sales
- Additionally, they can use information to construct a loyalty program so that buying from one item, ends with consumers getting points which can be redeemed towards other items