WALMART RECRUITING: TRIP TYPE CLASSIFICATION PROJECT

BY SIMRAN CELLINI

PRESENTATION OUTLINE

- Project motivation
- Project aim.
- Gathering and analyzing my data.
- Approach
- Data Exploration.
- Data Exploration Continued.
- Methodology.
- Decision Tree Results
- Gradient Boosted Trees Results.
- Support Vector Machine Results
- Conclusion

PROJECT MOTIVATION

When researching online and reading a customer survey I found out the following:

- Only 30% of the people believed that quick checkout is important.
- 50% said they cared more about prices.
- 70% preferred a convenient store location over anything else.

I chose this particular topic as there are lots of different trip types that people do, whether they are on a last minute run for pet food, leisurely making their way through a weekly grocery list or even just want to return one item this meaning that there will be a lot of data to be able to use and analyze with.

PROJECT AIM

BASED ON THE DIFFERENT SHOPPING TRIP FEATURES CAN WE PREDICT WHAT THE TRIP TYPE WILL BE?

GATHERING & ANALYSING MY DATA

The first step was to gather and analyze my data. The table below shows all of the 6 different types of shopping features and the output of the project which is the Trip Type. My table has a total 79 different types of house feature.

	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
5	26	8	Friday	2.006619e+09	2	PAINT AND ACCESSORIES	1017.0

The different features.

This is the output that I am trying to predict.

TRIP TYPE: A categorical ID representing the type of shopping trip the customer made.
VISIT NUMBER: An id corresponding to a single trip by a single customer.
WEEKDAY: The weekday of the trip.

UPC: The UPC number of the product purchased.
SCAN COUNT: The number of the given item that was purchased. A negative value indicates a product return.
DEPARTMENT DESCRIPTION: A description of the items department.
FINELINE NUMBER: A more refined category for each of the products.

APPROACH

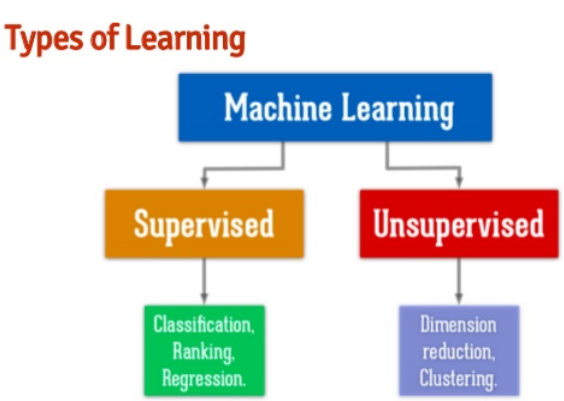
Inputs were taken from the table.

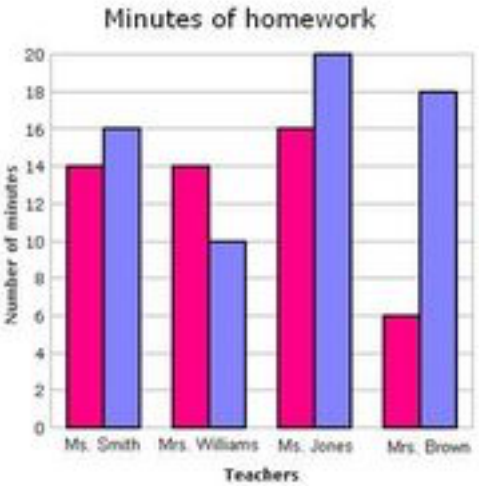
Model built on the training set to predict what the trip type will be.

Validate / Test the model on the test set . Explore data to understand which inputs are most important.

Analyse the results.

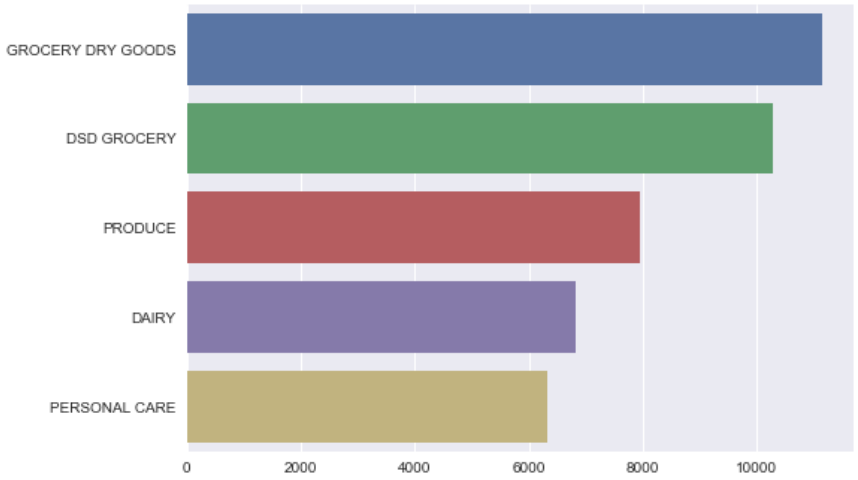
Weekday	Upc	ScanCount
Sunday	4.900004e+09	1
Friday	3.500039e+09	1
Tuesday	3.343620e+10	2
Saturday	8.437470e+10	1
Sunday	1.312000e+09	1



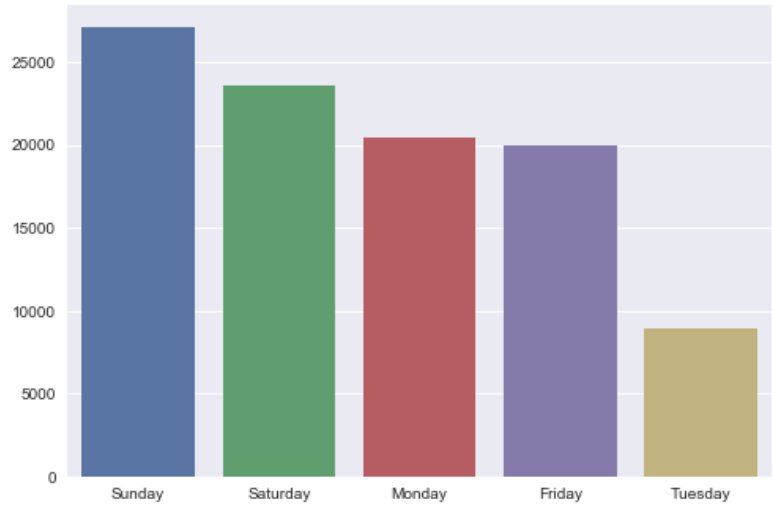


DATA EXPLORATION

In my data exploration I created a variety of different bar charts to show the distribution of the important variables. The first bar chart shows the most popular departments in Walmart. The second bar chart shows the most popular shopping day at Walmart.



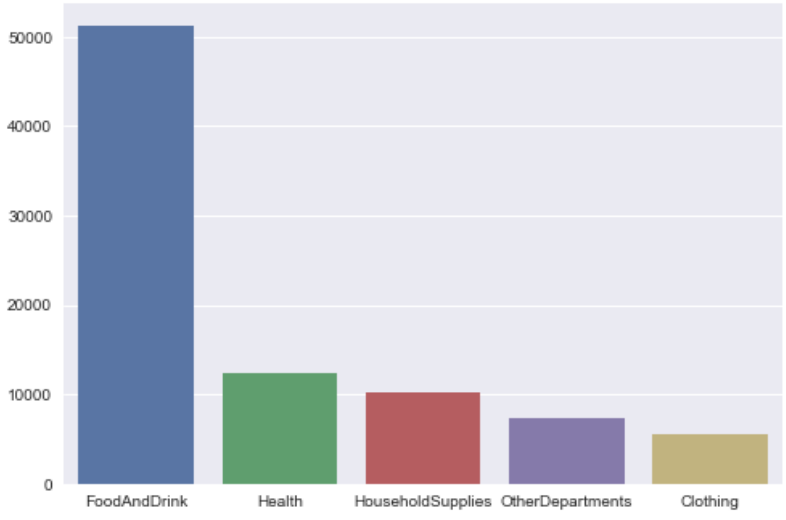
Most Popular Departments



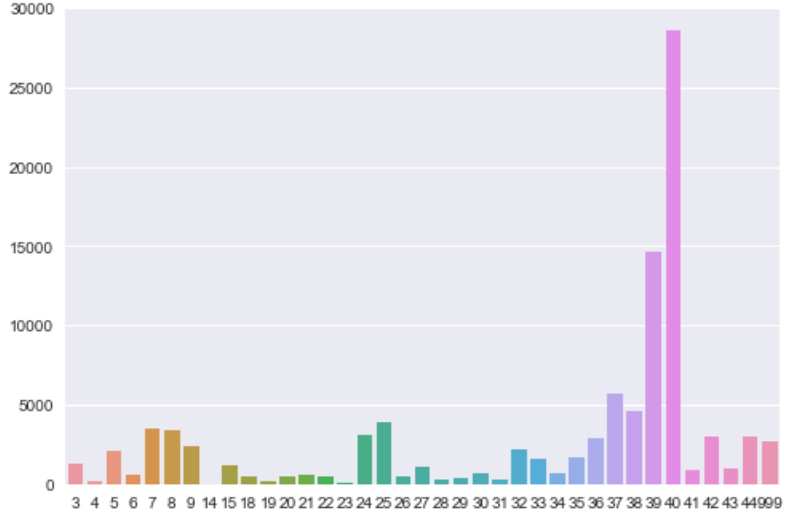
Most Popular Shopping Day

DATA EXPLORATION CONTINUED

For the third bar chart I grouped together the different departments into specific and more refined categories. The fourth bar chart shows the different types of Trip Types when customers shopped at Walmart. There are 37 different trip types.



Refined Store Categories



Popular Trip Types

METHODOLOGY

Below are the three machine learning models that I used.

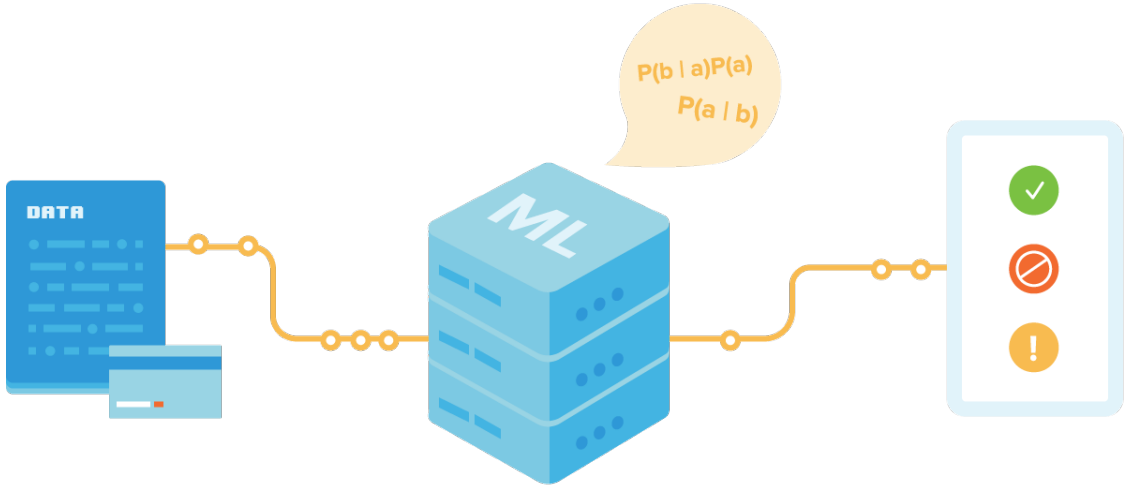
Models

Decision Trees (Makes predictions by choosing one input at a time splitting the data)
Gradient Boosted Trees(Made up of decision trees, each tree is a weak classifier, adds the outputs and then takes the average output.
Support Vector Machine (Splits the two classes by the inputs, the largest separation the better)

Evaluation

Classification Output

Accuracy
Classification Report



DECISION TREE RESULTS

The original Precision & Recall table shows all of the 37 different Trip Type results. I selected 5 of them.

PRECISION & RECALL

	Precision	Recall	F1 - Score	Support
3	0.85	1.00	0.92	226
8	0.58	0.77	0.66	542
9	0.58	0.63	0.60	403
40	0.68	0.74	0.71	293
999	0.97	0.81	0.89	374
Avg/Total	**0.51**	**0.53**	**0.51**	**4357**

ACCURACY

53%

GRADIENT BOOSTED TREES

PRECISION & RECALL

	Precision	Recall	F1 - Score	Support
8	0.58	0.77	0.66	542
9	0.58	0.63	0.60	403
25	0.43	0.62	0.51	150
39	0.39	0.51	0.44	468
999	0.97	0.81	0.89	374
Avg/Total	**0.51**	**0.53**	**0.51**	**4357**

ACCURACY

53%

SUPPORT VECTOR MACHINE RESULTS

PRECISION & RECALL

	Precision	Recall	F1 - Score	Support
8	0.12	0.01	0.02	542
24	0.05	0.01	0.01	129
39	0.41	0.29	0.34	468
40	0.70	0.65	0.67	293
999	0.38	0.63	0.48	374
Avg/Total	**0.24**	**0.25**	**0.20**	**4357**

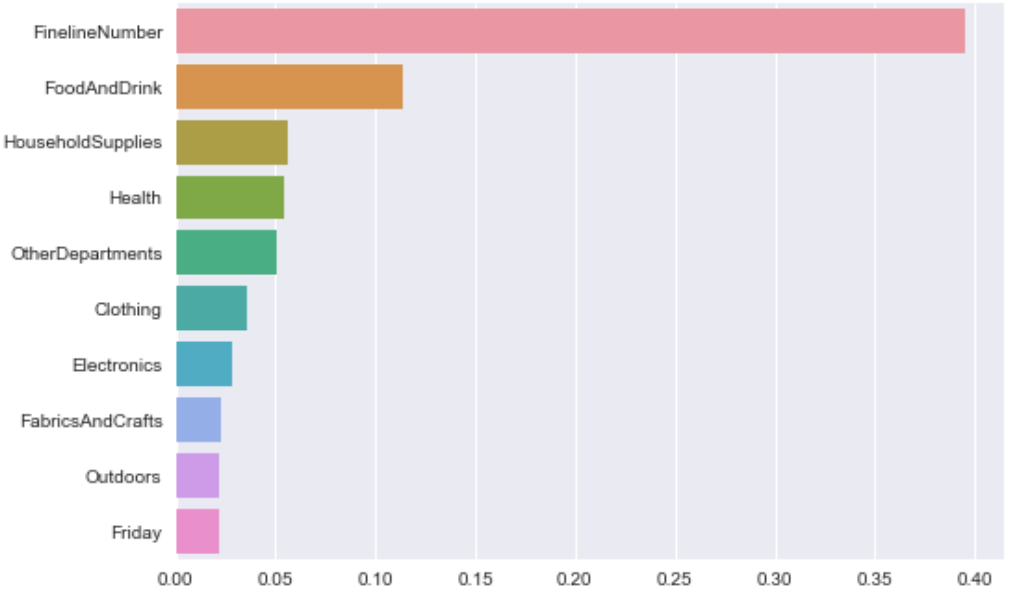
ACCURACY

25%

CONCLUSION

- To sum up the best models to use was the Decision Tree model and Gradient Boosted Trees. They both had the same accuracy of 53%.
- The most important part of this project was to build good predictor features of the Trip Type.

- I also learnt what the top ten important features are relating to the Trip Type. These are as follows:



ANY QUESTIONS ?