STEP
ƒUNCTION

WALMART TRIP TYPE CLASSIFICATION

-Ankita Sarawgi

Introduction

- Walmart: Multinational retail corporation ($485 billion annual revenue)
- Predicting customer trip types based on the items that the customer purchase
- Source: Kaggle Competitions
- Training set instances: (0.7 million)

Datasets

- Trip types examples: -Daily Dinner trip
  -Weekly grocery
  -Seasonal clothes
  -Holiday trips(Christmas)
- 38 Trip Types (Numbers and not what they represent)

Visit No.	• ID corresponding to single trip by a single customer
Weekday	• Weekday of the trip made
Upc	• Upc number of the product
Scan Count	• Number of items purchased
Department	• Department to which the item belonged
Fineline number	• High level refined category for product

Scope And Limitation

- Usage: -Product placement

  -Improving the shopping experience of the customers
- Challenges involved - Instances according to each item

  - Non descriptive data(Upc and Fineline Number)

  - Large dataset(>1.2 million)

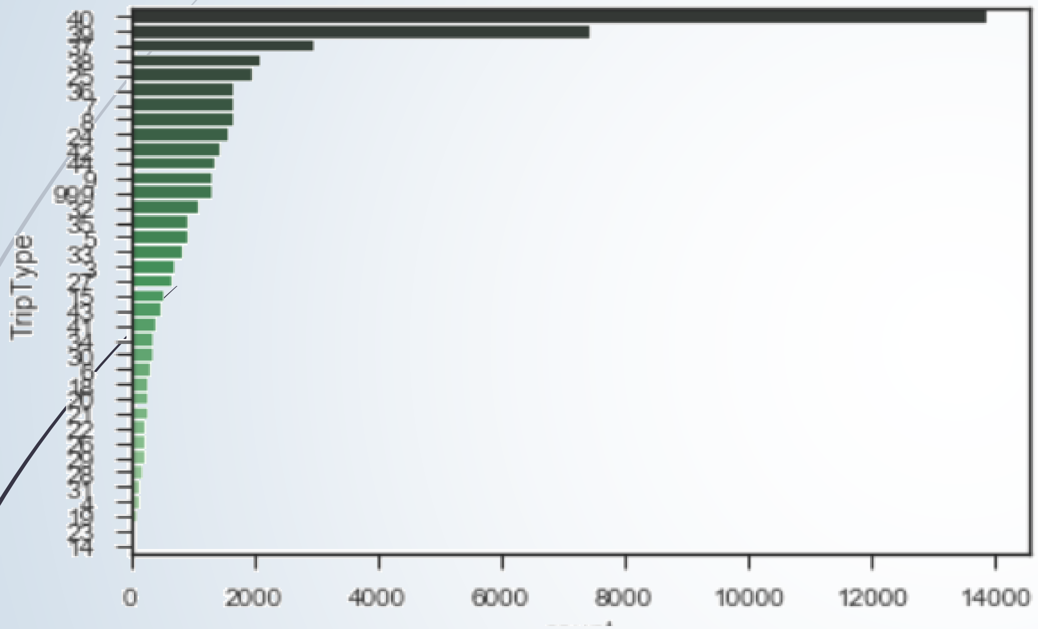
  - Powerful computer required

Feature Engineering

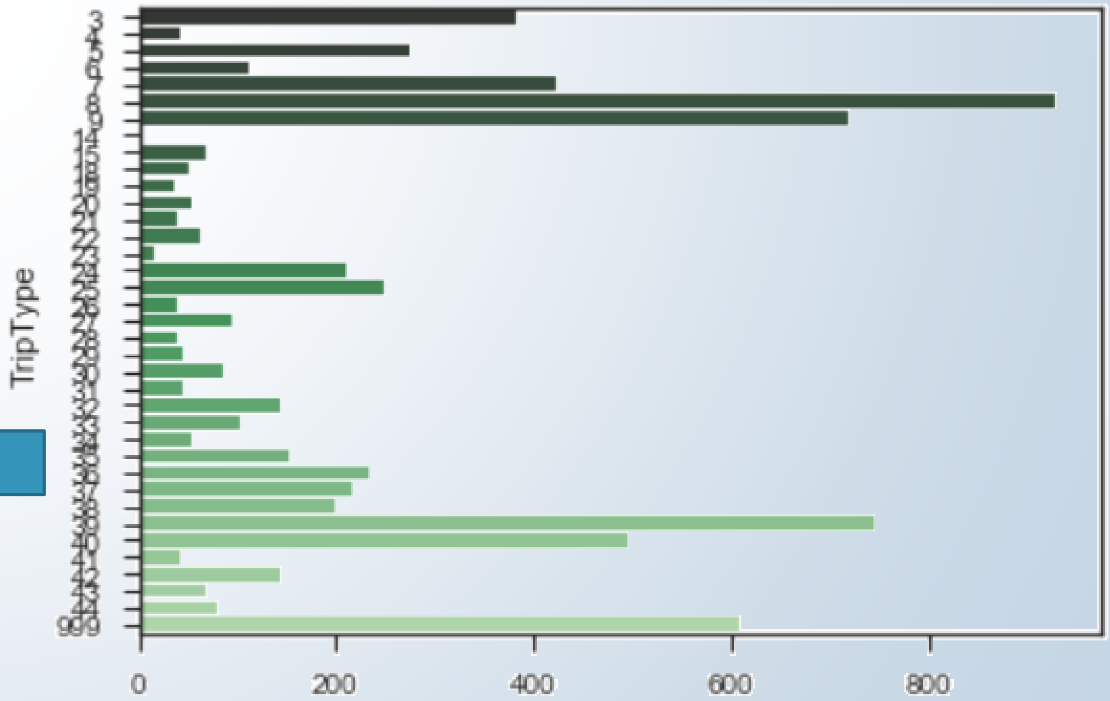
- “Success of all Machine Learning algorithms depends on how you present the data”
- It is manually transforming data into things our algorithm can understand
- Flexibility, simpler models and better results
- Achieved in two steps -Brainstorming and exploration

  -Creating new features

Data Exploration



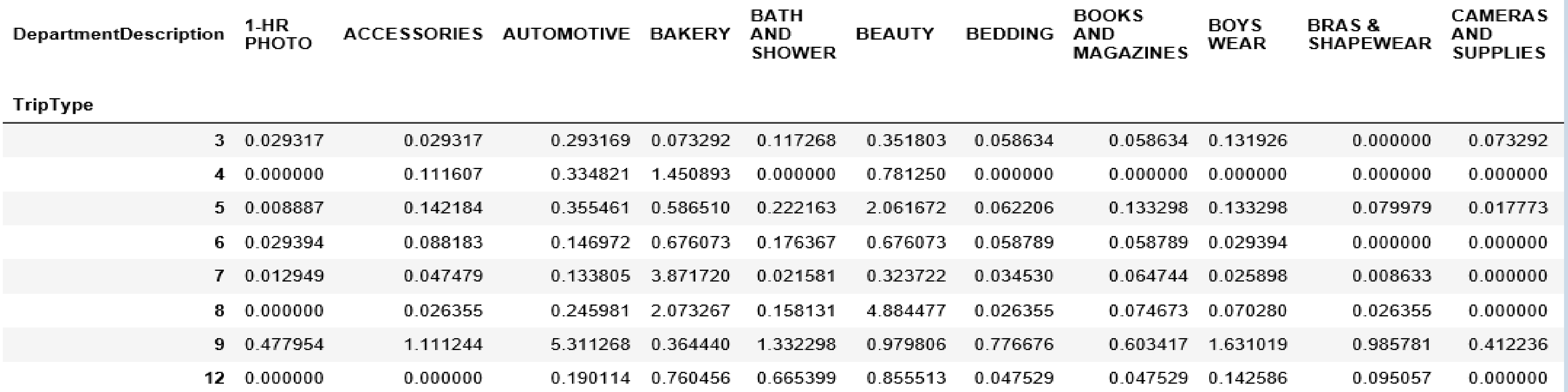
-Trip types and its frequency
-Number of different products in each type



-More efficient way of binning
-”Rare”, “Special”, “Medium”, “Frequent

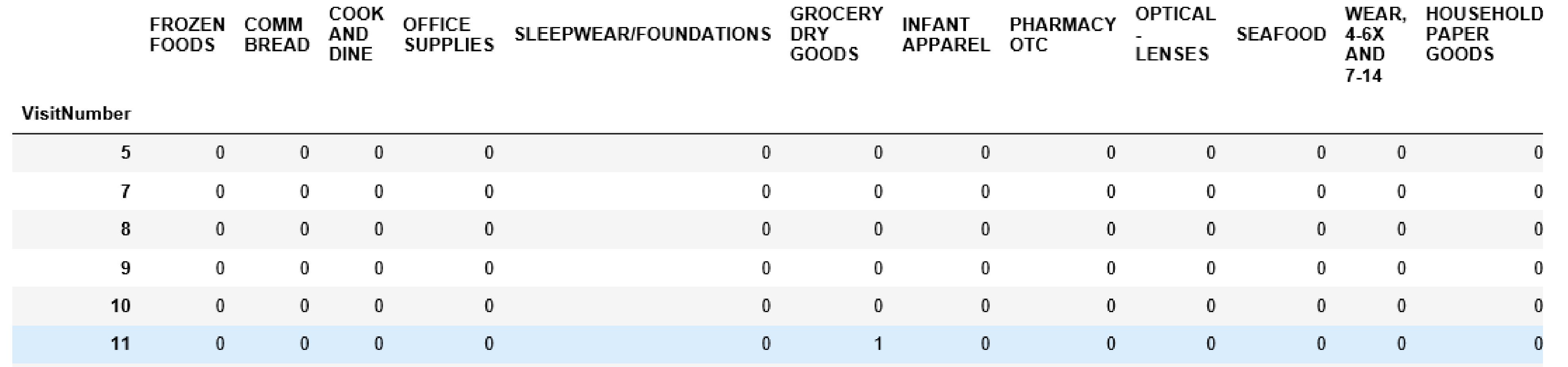
Data Exploration Continued…

Cross Tabulation



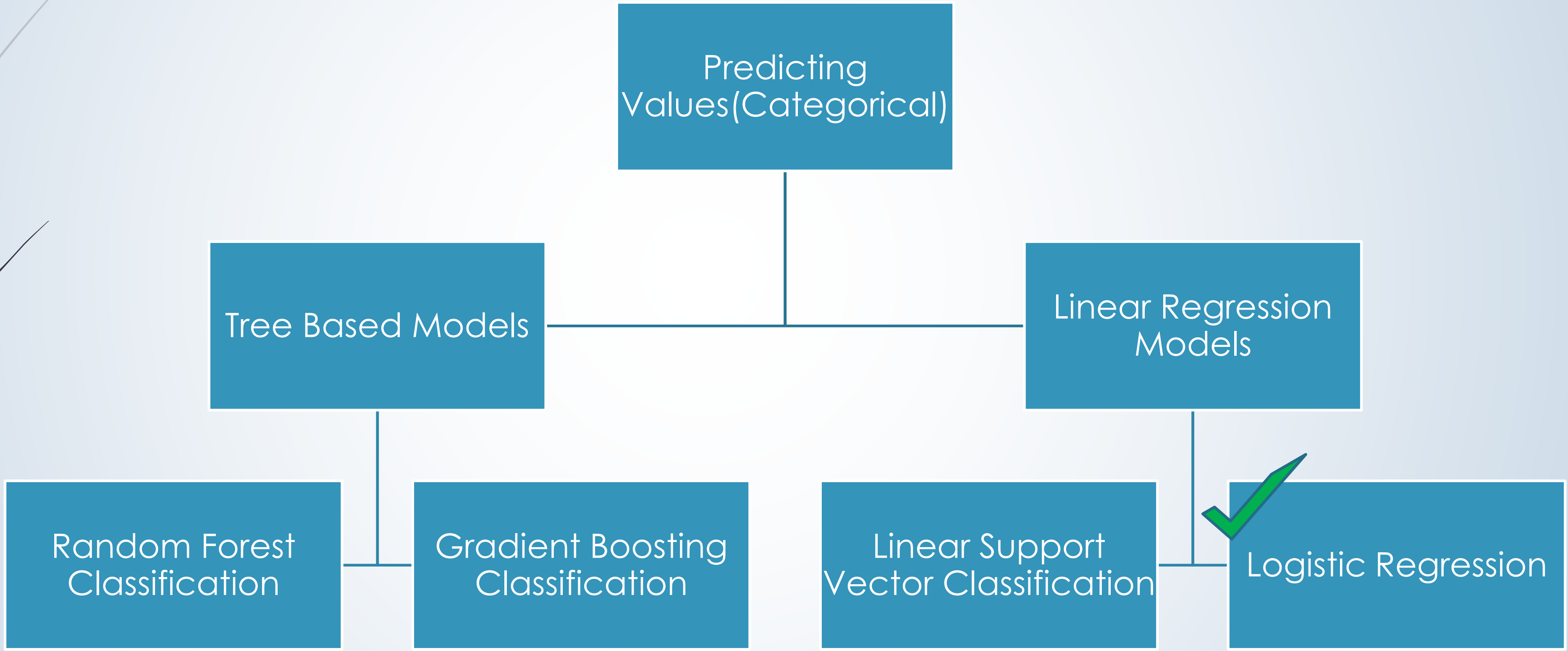
DepartmentDescription	1-HR PHOTO	ACCESSORIES	AUTOMOTIVE	BAKERY	BATH AND SHOWER	BEAUTY	BEDDING	BOOKS AND MAGAZINES	BOYS WEAR	BRAS & SHAPEWEAR	CAMERAS AND SUPPLIES
TripType											
3	0.029317	0.029317	0.293169	0.073292	0.117268	0.351803	0.058634	0.058634	0.131926	0.000000	0.073292
4	0.000000	0.111607	0.334821	1.450893	0.000000	0.781250	0.000000	0.000000	0.000000	0.000000	0.000000
5	0.008887	0.142184	0.355461	0.586510	0.222163	2.061672	0.062206	0.133298	0.133298	0.079979	0.017773
6	0.029394	0.088183	0.146972	0.676073	0.176367	0.676073	0.058789	0.058789	0.029394	0.000000	0.000000
7	0.012949	0.047479	0.133805	3.871720	0.021581	0.323722	0.034530	0.064744	0.025898	0.008633	0.000000
8	0.000000	0.026355	0.245981	2.073267	0.158131	4.884477	0.026355	0.074673	0.070280	0.026355	0.000000
9	0.477954	1.111244	5.311268	0.364440	1.332298	0.979806	0.776676	0.603417	1.631019	0.985781	0.412236
12	0.000000	0.000000	0.190114	0.760456	0.665399	0.855513	0.047529	0.047529	0.142586	0.095057	0.000000

New set of features



	FROZEN FOODS	COMM BREAD	COOK AND DINE	OFFICE SUPPLIES	SLEEPWEAR/FOUNDATIONS	GROCERY DRY GOODS	INFANT APPAREL	PHARMACY OTC	OPTICAL - LENSES	SEAFOOD	WEAR, 4-6X AND 7-14	HOUSEHOLD PAPER GOODS
VisitNumber												
5	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	1	0	0	0	0	0	0

Datasets and Modelling



Predicting Values(Categorical)

Tree Based Models

Linear Regression Models

Random Forest Classification

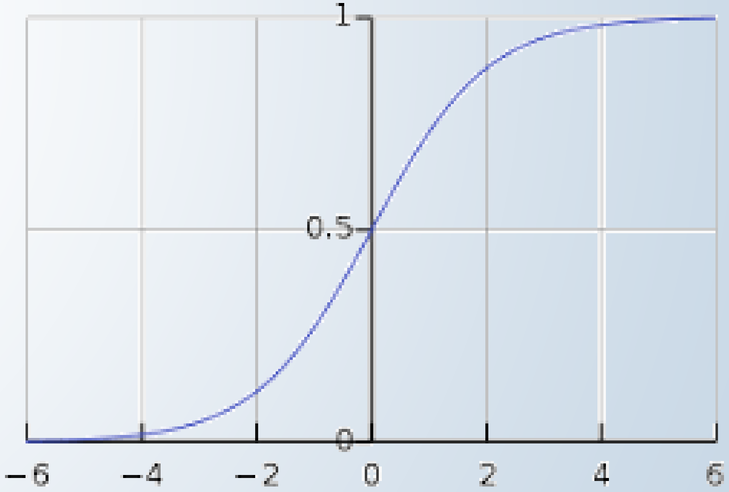
Gradient Boosting Classification

Linear Support Vector Classification

✓ Logistic Regression

Logistic Regression

- Key representation in logistic regression are the coefficients, just like linear regression
- Coefficients in logistic regression are estimated using a process called maximum-likelihood estimation
- Probability using sigmoid function
- Conducted on one class against all others
- Repeated such that all classes are regressed



Results

- Different evaluation method used for the problem like accuracy score, classification and confusion matrix

	Precision	Recall	F1 Score
Random Forest classification	.59	.45	.41
Gradient Boosting Classification	.60	.59	.58
Linear Support Vector	.56	.57	.55
Logistic Regression	.58	.59	.57

Conclusion

- Biggest Challenge: Limitation of solving large data

  -Better performance by considering more training points

  -More features to be considered

- Feature Engineering was the key