Walmart Trip Type Classification

Group 5

Gokul Gunasekaran
Monish Chandrasekhar
Vhinesh Ravi

Motivation

- Refine Walmart’s segmentation process.
- Improve the science behind trip type classification.
- Enhance product placement and product assortment.
- Understand what type of products customers buy and improve their shopping experience.
- Understand customers shopping pattern over the week.

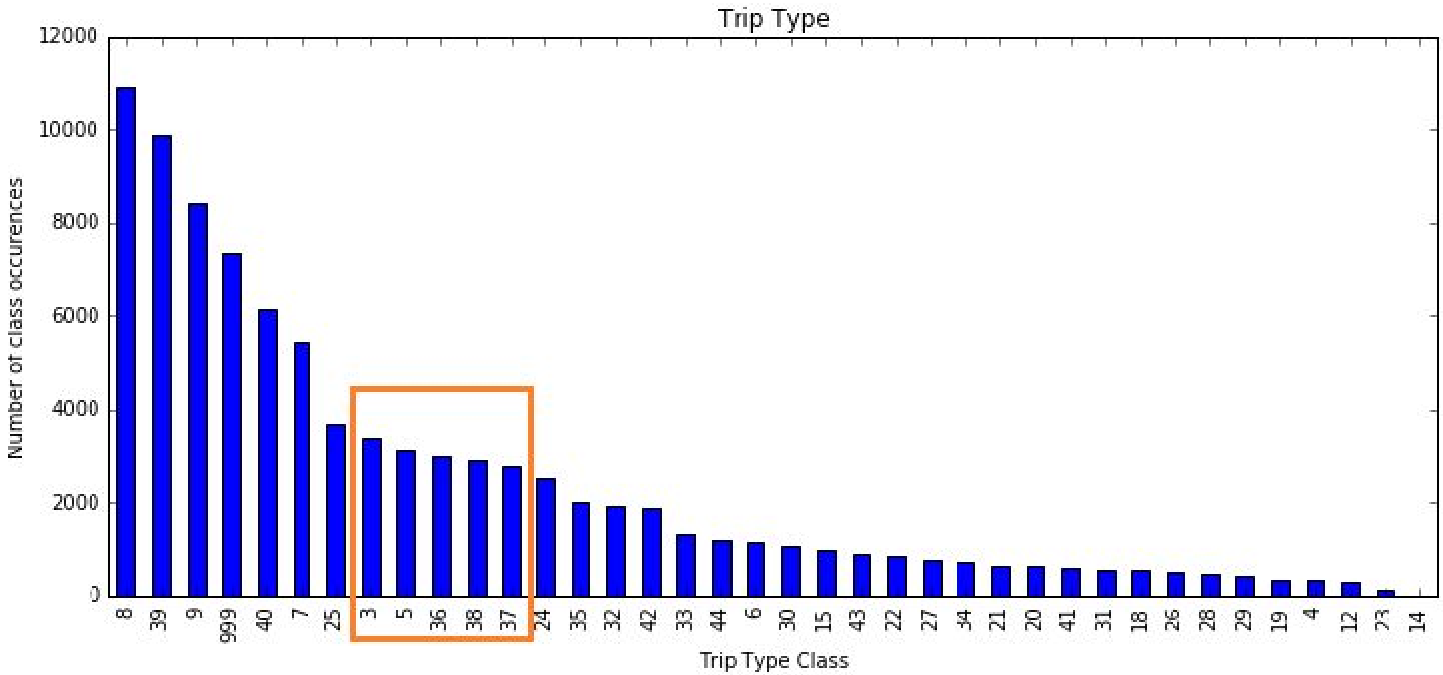
Problem Definition

- Improve Walmart’s segmentation process by classifying the customer trip types based on the product purchased in that visit, so that it creates the best shopping experience for every customer.
- Solutions
    - Manual
    - Rule based
    - Machine learning

Dataset information

- Kaggle competition
- Number of records - 647,054 with 38 trip types
- Data fields
  - TripType - Type of shopping trip
  - Visit Number - ID of the trip
  - Weekday
  - UPC - unique number of the product
  - Scan Count - Number of items purchased. Product return is a negative number
  - Department Description - Description of item’s department
  - Fineline Number - Refined category of each product

Class Distribution



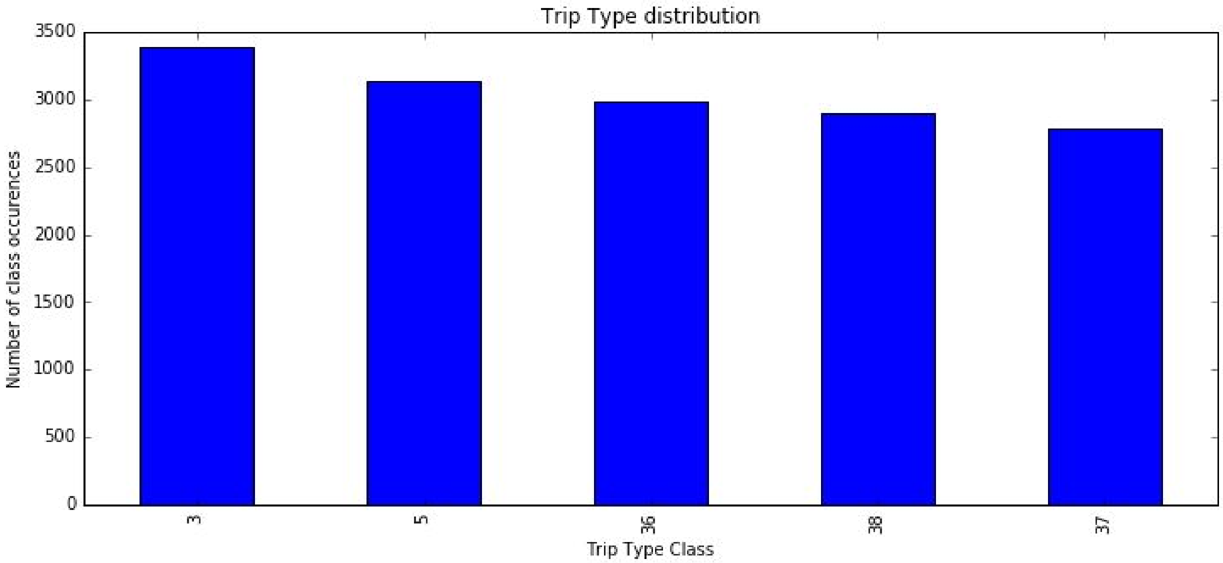
Subset

- Used a subset of the data. i.e. Number of instances we used ~90k with 5 classes.
- Data Snapshot

	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

Class Distribution

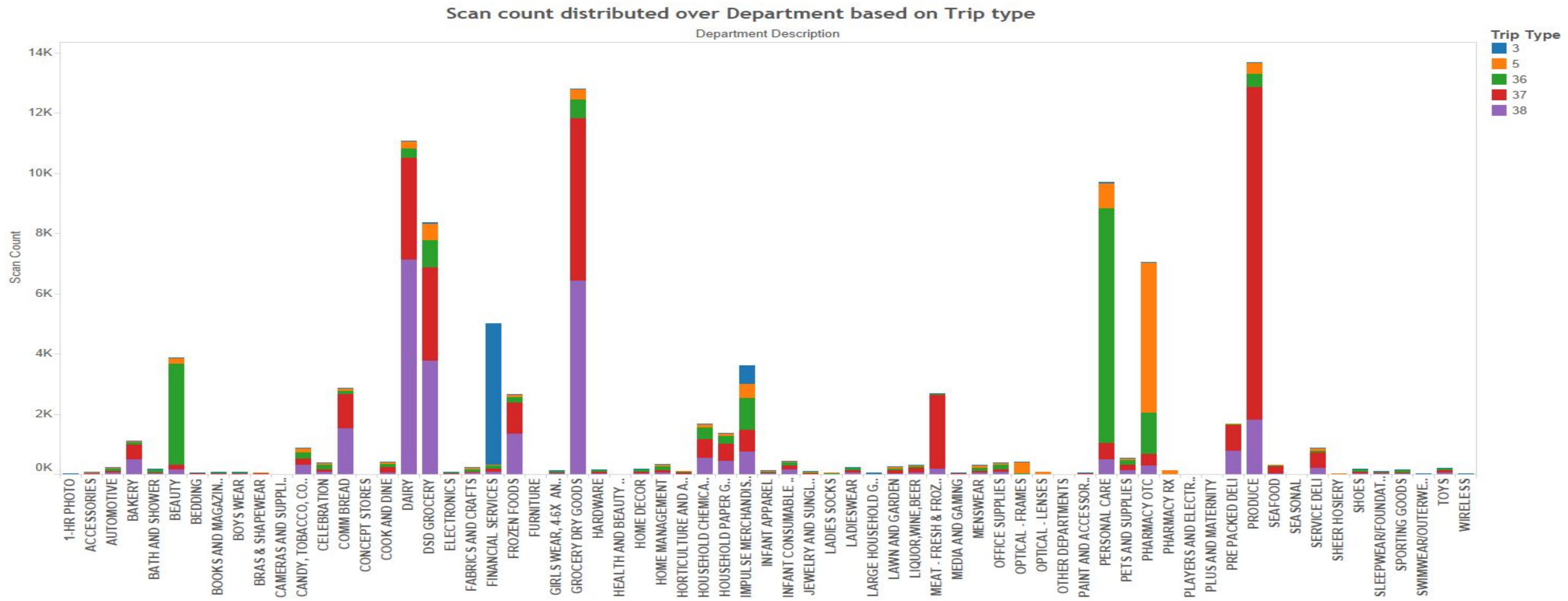
- The trip type against number of class occurrences revealed that Class 3 was the most frequent trip type .



Challenges

- Each record was representing an item instead of a visit.
    - We grouped the records by their visit number to classify the trip.
- Records with missing values ~4000 rows
    - We removed the records with NULL, Blank values
- Dummy variables(Categorical) - Weekday
    - Converted qualitative values to quantitative
    - Eg: Monday =1, Tuesday =2
- Duplicate department labels
    - We identified and combined them into a single category
    - Eg: “MENSWEAR” and“MENS WEAR”

EDA - TripType vs Department vs ScanCount



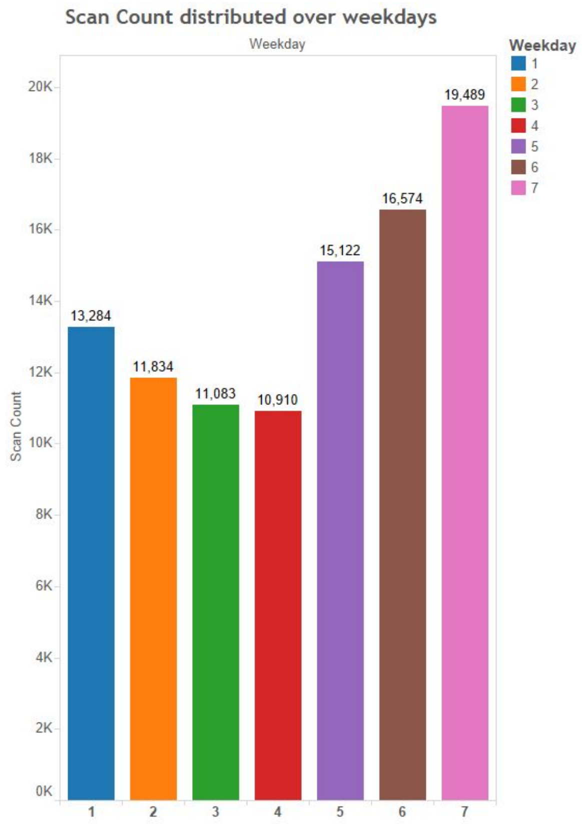
Sum of Scan Count for each Department Description. Color shows details about Trip Type. The view is filtered on Trip Type, which keeps 3, 5, 36, 37 and 38.

Inferences and Assumptions

- Trip Type Class 3 - Financial Services, impulse merchandise
- Trip Type Class 5 - Pharmacy, Personal care
- Trip Type Class 36 - Personal care, beauty, pharmacy
- Trip Type Class 37 - Produce,grocery dry,Meat, Dairy
- Trip Type Class 38 - Dairy, bread, breakfast foods,grocery dry

EDA - Weekday against Scan Count

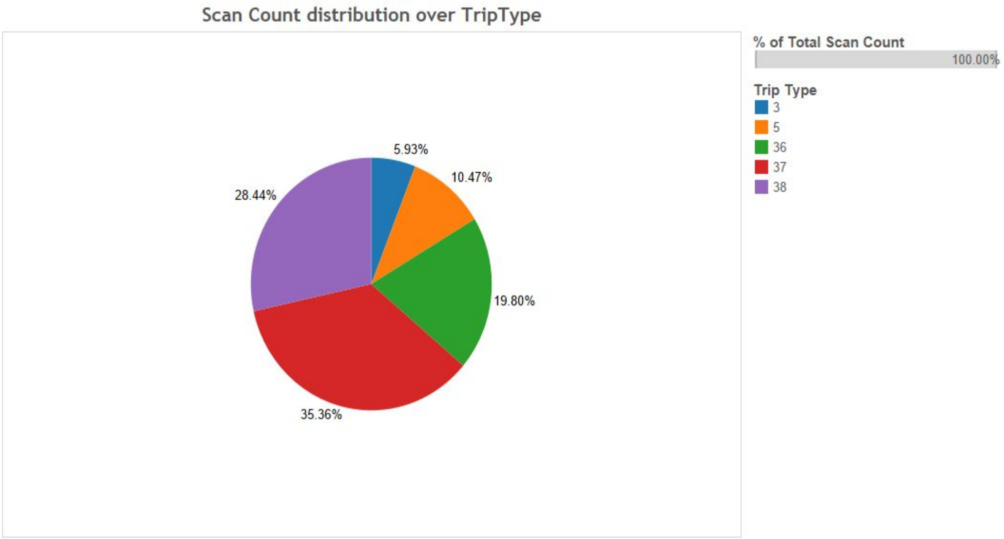
- More products are sold on Sundays.
- Very few products are sold on Thursdays.



Sum of Scan Count for each Weekday. Color shows details about Weekday.

EDA - TripType against ScanCount

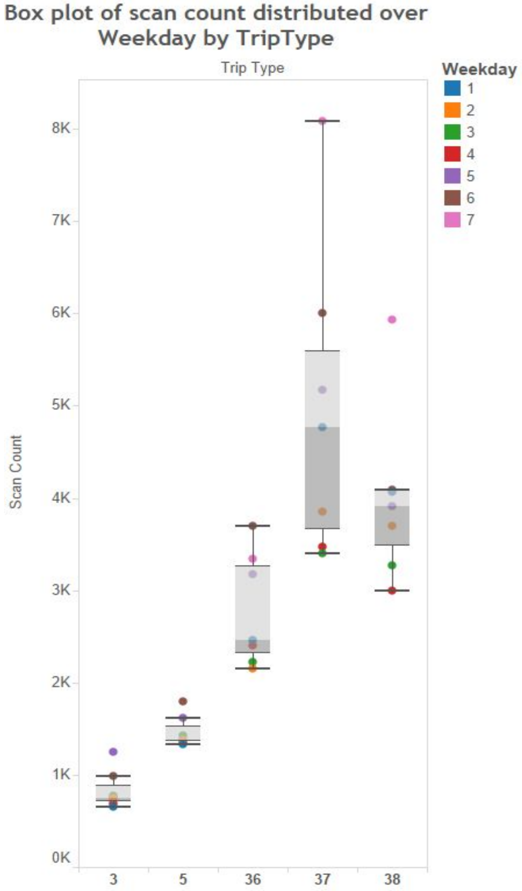
- Products from Class 37(Produce, Grocery dry, Meat, Dairy) are sold on a larger scale compared to other classes.



Trip Type (color) and % of Total Scan Count (size). The view is filtered on Trip Type, which keeps 3, 5, 36, 37 and 38. Percents are based on each row of the table.

EDA - TripType VS WeekDay vs ScanCount

- Class 3(Financial Services) is mostly used on Fridays.
- Other classes have more products sold on Weekends.

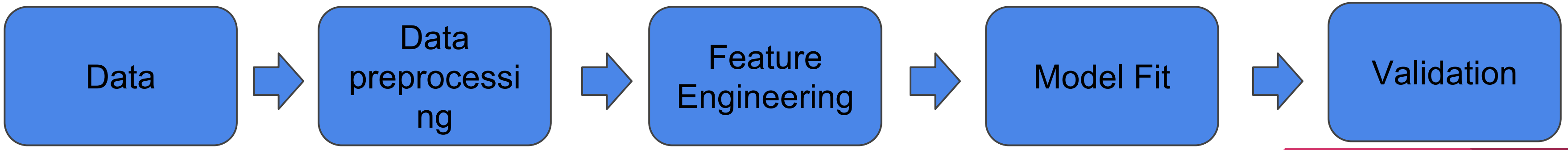


Sum of Scan Count for each Trip Type. Color shows details about Weekday. The view is filtered on Weekday, which keeps 7 of 7 members.

Related work

Approach

- Initial model were Decision Trees (40%), KNN(61%), Random forest(54.77%).
- Analysis produced low accuracies for all three models.



Feature Engineering

- Multidimensional data (several rows per customer visit).
- From the data visualization, the trip type is more dependent on department description. Hence feature engineering from long to wide format.
- From 7 dimensions to 71 dimensions.

Feature Engineering

- Added individual departments as additional columns.
- Aggregated total number of products bought on each visit.
- Aggregated total number of products bought in each department on each visit.
- Added a Return column representing product returns.

Dimensions of resulting data: 15195 x 71

VisitNumber	TripType	Weekday	NumItems	Return	1-HR PHOTO	ACCESSORIES	AUTOMOTIVE	BAKERY	BATH AND SHOWER	...	SEAFOOD	SEASONAL	SE DI
43	38	5	4	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
63	36	5	5	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
83	36	5	9	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
86	37	5	22	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
97	38	5	13	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.

Random Forest

- Ensemble classifier with 100 estimators. Training accuracy - 98%, Testing accuracy - 88%

Confusion Matrix for Random Forest:

Random Forest	N = 100						
		Predicted					
		3	5	36	37	38	All
TRUE	3	1021	8	6	1	3	1039
	5	12	789	80	19	22	922
	36	9	78	769	15	24	895
	37	1	12	13	768	71	865
	38	5	25	27	90	691	838
	All	1048	912	895	893	811	4559

Classification Report

	precision	recall	f1-score	support
3.0	0.97	0.98	0.97	1007
5.0	0.85	0.86	0.85	912
36.0	0.86	0.86	0.86	928
37.0	0.87	0.88	0.88	833
38.0	0.86	0.84	0.85	879
avg / total	0.88	0.88	0.88	4559

KNN

- Non parametric classifier (k = 5)
- Training accuracy = 89% , Testing accuracy = 86%

Confusion Matrix for KNN:

KNN	K = 5						
		Predicted					
		3	**5**	**36**	**37**	**38**	**All**
TRUE	**3**	1019	14	3	0	3	1039
	5	24	791	62	14	31	922
	36	13	94	746	15	27	895
	37	4	20	13	701	127	865
	38	7	36	28	104	663	838
	All	1067	955	852	834	851	4559

Classification Report

	precision	recall	f1-score	support
3	0.94	0.98	0.96	1007
5	0.84	0.84	0.84	912
36	0.86	0.85	0.85	928
37	0.85	0.81	0.83	833
38	0.80	0.81	0.81	879
avg / total	0.86	0.86	0.86	4559

Linear Discriminant Analysis

- Linear Classifier
- Training Accuracy = 86% , Testing accuracy = 86%

Confusion Matrix for LDA:

LDA							
		Predicted					
		3	5	36	37	38	All
TRUE	3	947	78	12	0	2	1039
	5	4	845	49	6	18	922
	36	19	100	732	13	31	895
	37	7	50	18	653	137	865
	38	10	45	22	35	726	838
	All	987	1118	833	707	914	4559

Classification Report

	precision	recall	f1-score	support
3.0	0.96	0.91	0.93	1007
5.0	0.74	0.91	0.82	912
36.0	0.90	0.84	0.87	928
37.0	0.94	0.76	0.84	833
38.0	0.82	0.87	0.84	879
avg / total	0.87	0.86	0.86	4559

Support Vector Machine

- Non-probabilistic binary linear classifier
- Training accuracy = 90%, Testing accuracy = 89%

Predicted	3.0	5.0	36.0	37.0	38.0	All
True						
3.0	1006	16	3	1	1	1027
5.0	10	825	48	14	37	934
36.0	4	80	767	17	26	894
37.0	0	13	12	731	82	838
38.0	2	31	22	66	741	862
All	1022	965	852	829	887	4555

	precision	recall	f1-score	support
3	0.97	0.97	0.97	988
5	0.85	0.88	0.87	961
36	0.89	0.85	0.87	920
37	0.89	0.85	0.87	831
38	0.84	0.88	0.86	859
avg / total	0.89	0.89	0.89	4559

Logistic Regression

- Measure the relationship between categorical dependent variable and independent variables by estimating probabilities.
- Training accuracy = 91%, Testing accuracy = 90%

Confusion Matrix for Logistic Regression:

Logisitc Regression

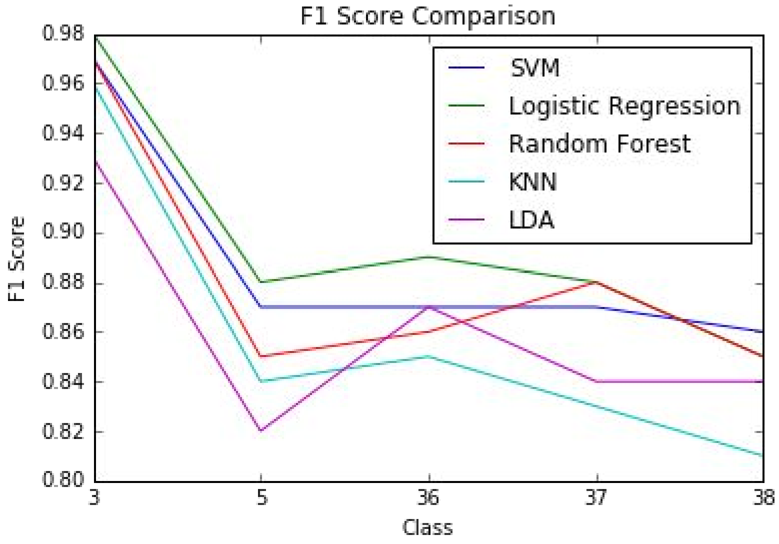
		Predicted					
		3	**5**	**36**	**37**	**38**	**All**
TRUE	**3**	1019	14	4	1	1	1039
	5	9	809	65	12	27	922
	36	12	67	781	14	21	895
	37	1	10	15	749	90	865
	38	2	28	21	61	726	838
	All	1043	928	886	837	865	4559

Classification Report

	precision	recall	f1-score	support
3.0	0.98	0.98	0.98	1007
5.0	0.87	0.88	0.88	912
36.0	0.89	0.89	0.89	928
37.0	0.89	0.87	0.88	833
38.0	0.84	0.86	0.85	879
avg / total	0.90	0.90	0.90	4559

Comparison

- Performance metric - F1 Score.
- Logistic Regression performs well.



Fine tuning Techniques

- 71 features can be reduced and sparsity can be handled using fine tuning.
- **PCA** - Experimented with different number of components: 1, 5, 10, 25, 30,70, yielding the same accuracy for both train and test sets.
- **Recursive Feature Elimination** - Experimented ranking different number of features: 10, 20, 30, 40, 60 and found that models produced a very low testing and training accuracy - underfitting
- **L1 Regularization** - Performed Cross-Validated Logistic Regression to get best (C) lambda [1,1,74,15,1] for classes[3,5,36,37,38]. This selected 63 important features for all classes.

L1 Regularization Logistic Regression

- Training Accuracy = 90%, Testing Accuracy = 90 %

Logisitc Regression L1 Regularization							
		Predicted					
		3	5	36	37	38	All
TRUE	3	1018	18	6	1	1	1044
	5	15	844	47	23	24	953
	36	8	72	792	17	22	911
	37	1	9	11	701	79	801
	38	3	21	27	65	734	850
	All	1045	964	883	807	860	4559

	precision	recall	f1-score	support
3.0	0.97	0.98	0.97	1044
5.0	0.88	0.89	0.88	953
36.0	0.90	0.87	0.88	911
37.0	0.87	0.88	0.87	801
38.0	0.85	0.86	0.86	850
avg / total	0.90	0.90	0.90	4559

Conclusion

- Transforming the raw data into features influenced the accuracy of the model on unseen data.
- Feature engineering and L1 Regularization helped fine tune the Logistic Regression model.
- PCA and Recursive Feature Selection failed to improve the selected models.

Future Work

- Based on the items purchased by the customer, generate personalised promotions and offers.
- Include all classes.

Thank you!