Market Basket Analysis

CS5661 SPRING 2018

MANIKANDAN ESWARAN

VIVEK AGARWAL

HARMINDER SINGH

Objective

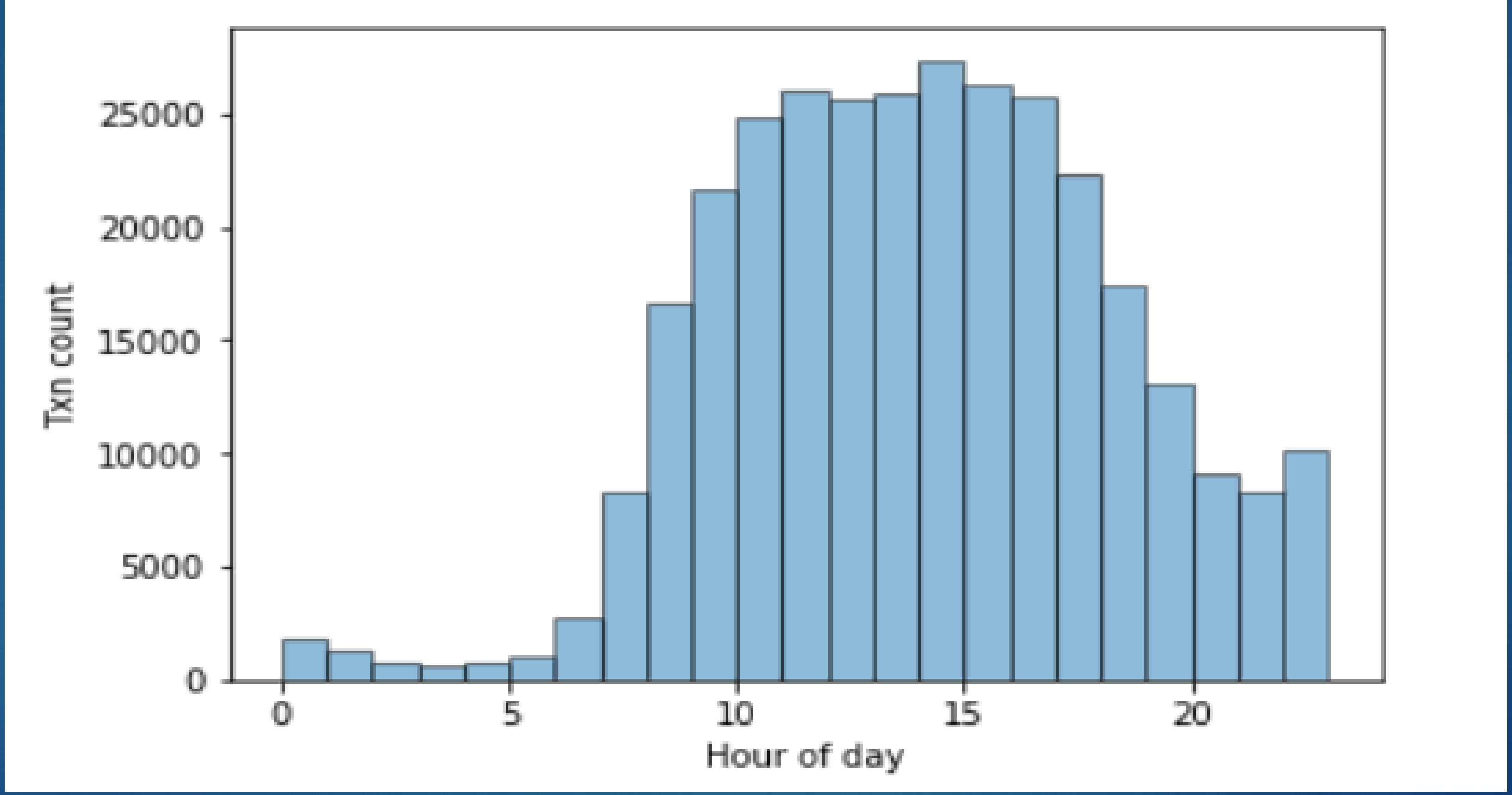
- Using standard scikit-learn algorithms and predict if a particular item (product) has the possibility to be re-ordered.
- Apply Apriori association rule data mining techniques to find the combination of product items that has a possibility of being ordered together with a given confidence level.

Dataset

- The dataset for this project was taken from kaggle. The data set contains transactions data set that contains a set of orders and the list of products in each order with a flag specifying if a particular item in that order was a re-order or the first order.

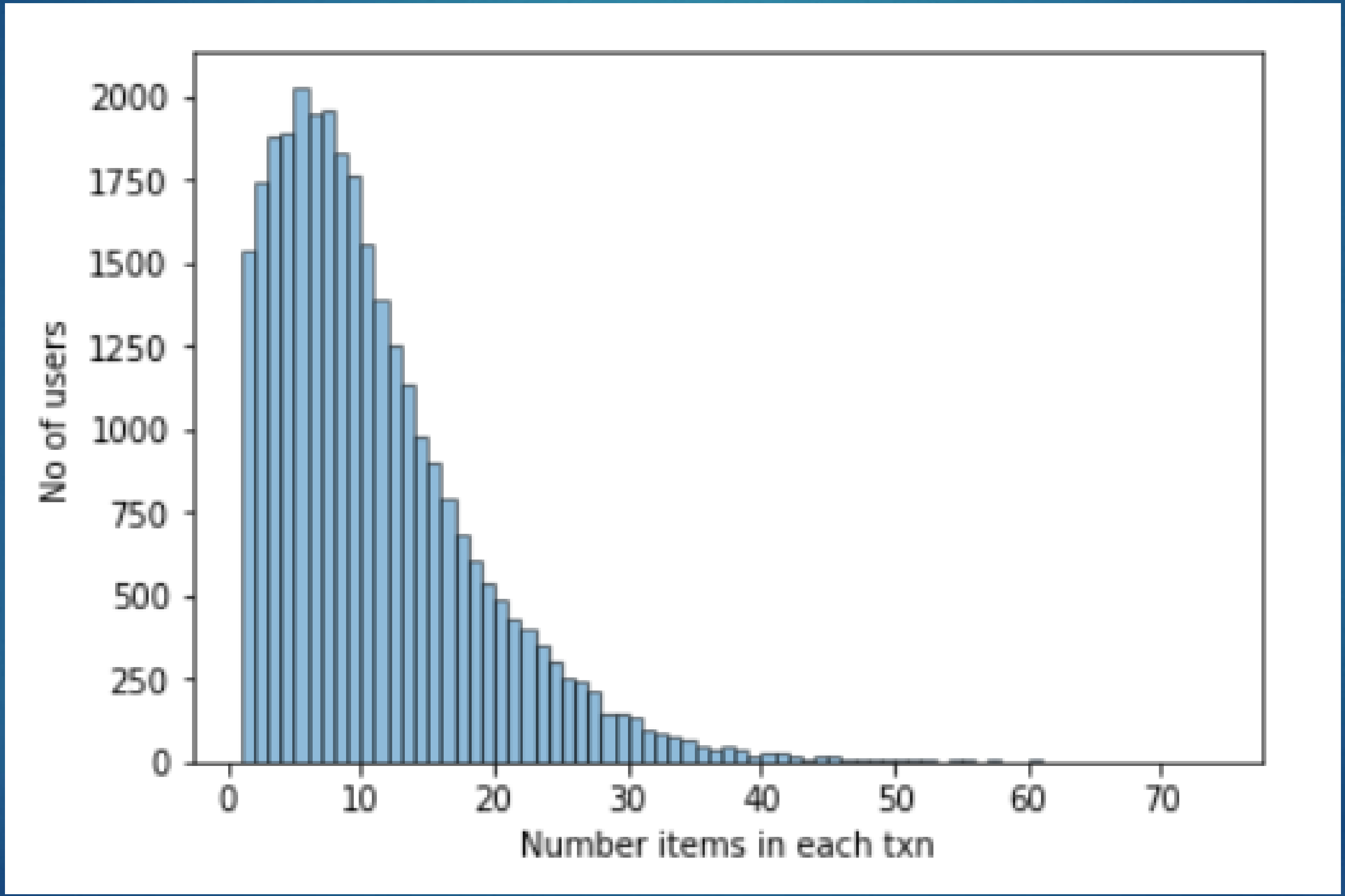
Understanding the dataset

Distribution of transactions over the day



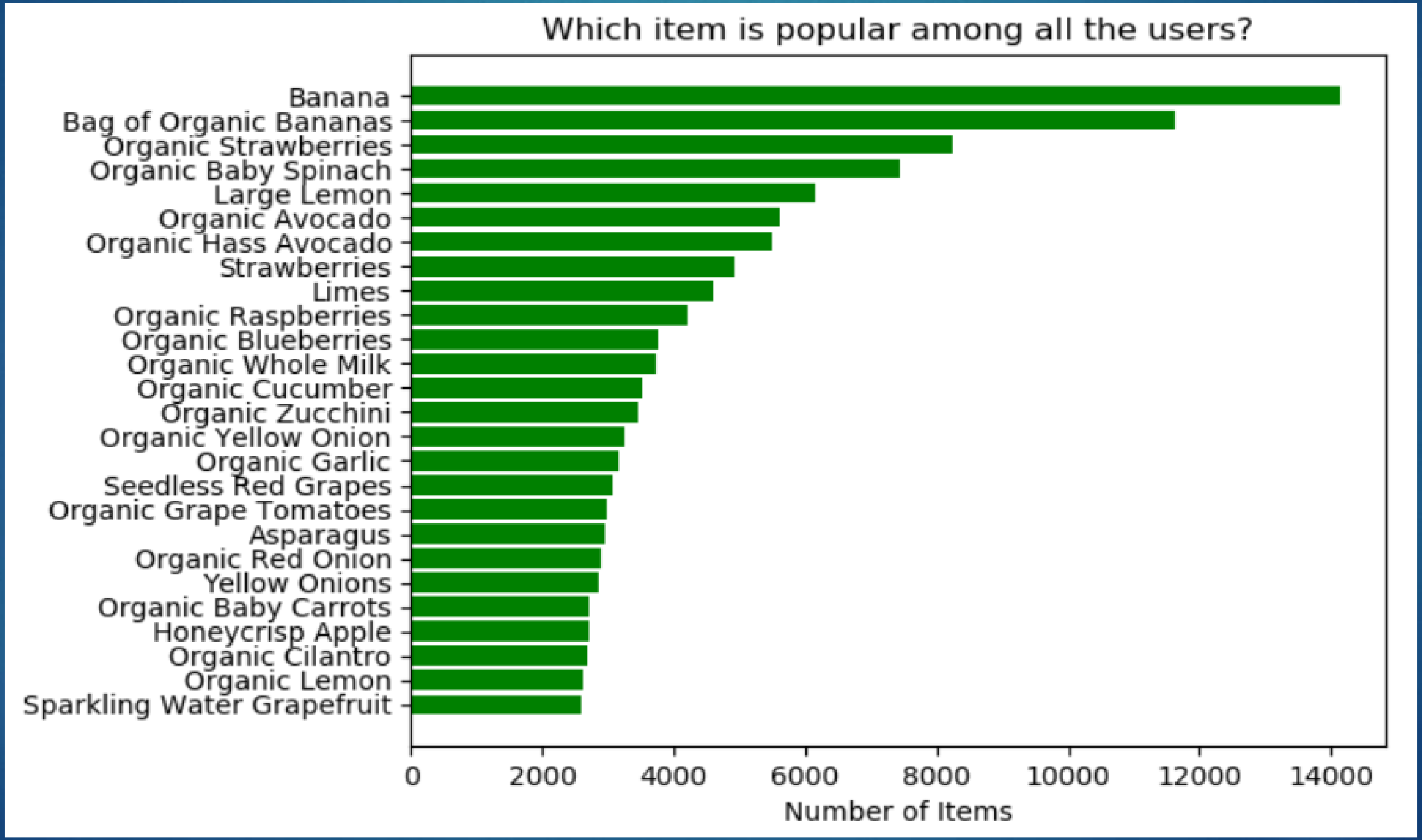
Understanding the dataset

Number of products purchased Vs number of users



Understanding the dataset

Most purchased products by users



Data preprocessing

- User id and product id values one-hot encoded. This resulted in 26k+ features.
- Applied PCA to reduce features to 50

Predicting Re-order flag

Algorithm	Parameters	Accuracy Score (%)
DecisionTreeClassifier	-	59.63
RandomForestClassifier	n_estimators=25	68.21

MLPClassifier	hidden_layer_sizes=(3,3)
	activation= 'logistic'
	solver='adam'
	alpha=1e-5, learning_rate_init = 0.01
		63.81

GridSearchCV/ MLPClassifier	{'hidden_layer_sizes': (5, 5)}	63.56
Deep NN	Dense(10/relu)->Dense(5/relu)->Dense(1/softmax)	60.45

Apriori data mining to find association rules

Transaction Id	Products
1	{milk, egg, bread}
2	{milk, egg, coffee, bread}
3	{sugar, coffee, toothbrush}
4	{milk, bread, coffee}
5	{sugar, egg, vinegar}

Product(s)	No of txns containing product(s)
{milk}	3
{bread}	3
{egg}	3
{sugar}	2
{toothbrush}	1
{milk, bread}	3
{egg, bread}	2
{sugar, toothbrush}	1

Apriori rules

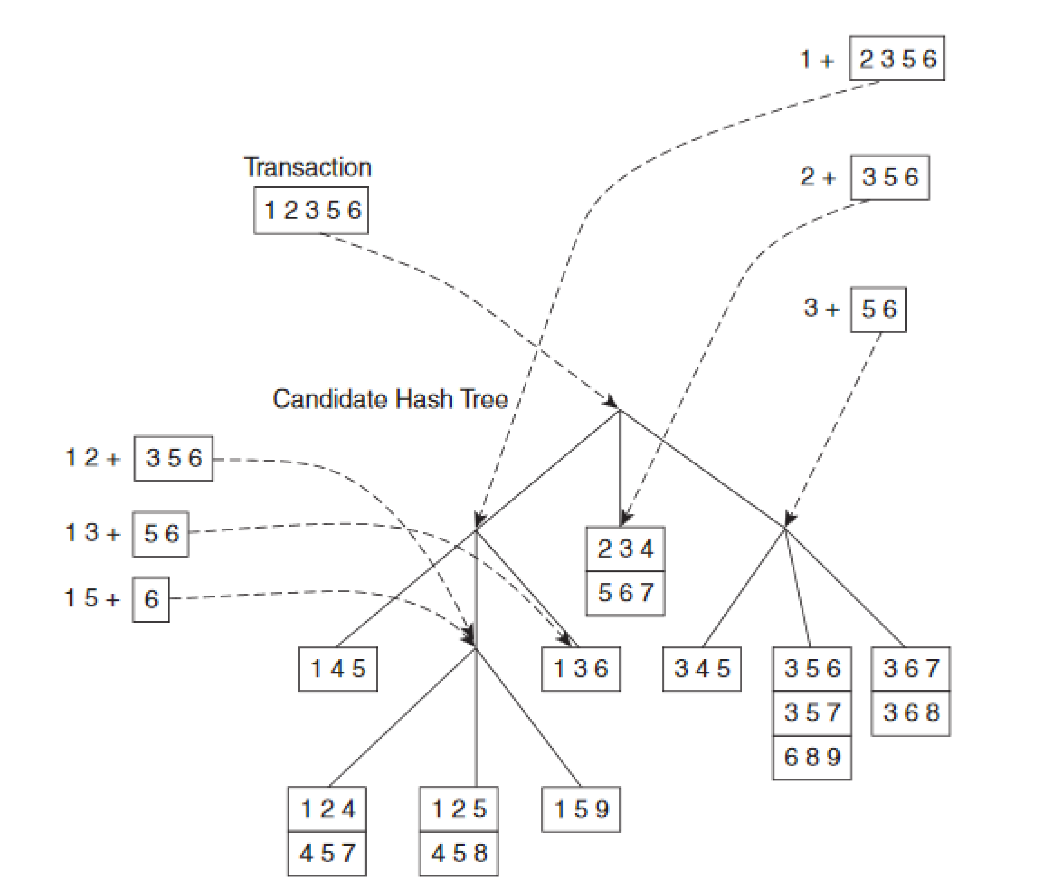
Rule	Support	Confidence
{milk}->{bread}	S({milk, bread})/N = 3/5 = 0.6	S({milk, bread})/S({milk}) = 3/3 = 1.0
{egg}->{bread}	S({egg, bread})/N = 2/5 = 0.4	S({egg, bread})/S({egg}) = 2/3 = 0.67
{sugar}->{toothbrush}	S({sugar, toothbrush})/N = 1/5 = 0.2	S({sugar, toothbrush})/S({sugar}) = 1/2 = 0.5

Support and Confidence

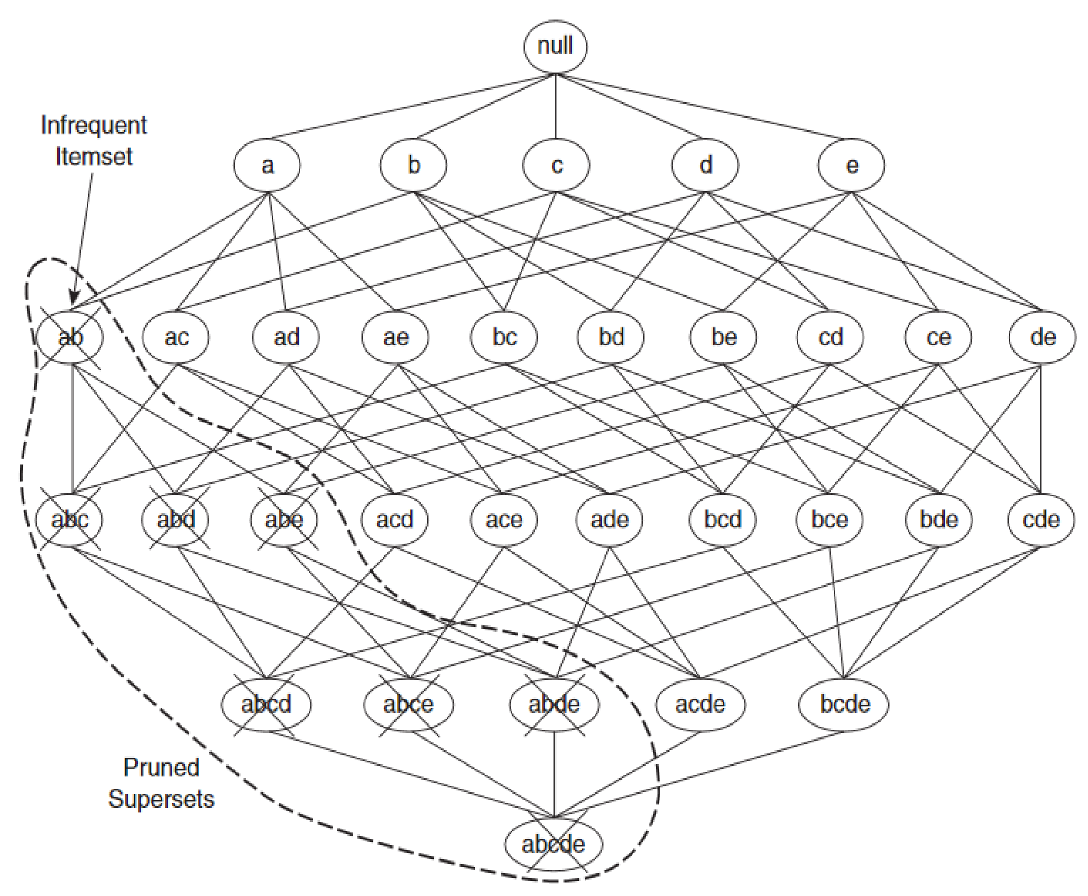
- **Support** – Support for rule {X}->{Y} is the ratio of number of transactions where a product group {X U Y} is part of, to the total number of transactions.
- $support(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$
- **Confidence** – Confidence of a rule {X}->{Y} is the ratio of support of the given product group {X U Y} to the support of the product group {X}.
- $confidence(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

Apriori Optimizations

Candidate set generation



Frequent item set generation



Apriori Results (50k Records)

Rule	Confidence
{Limes,Bunched Cilantro} -> {Large Lemon}	0.50
{Organic Red Bell Pepper,Banana} -> {Organic Avocado}	0.58
{Broccoli Crown,Organic Strawberries} -> {Banana}	0.55
{Seedless Red Grapes,Organic Baby Spinach} -> {Banana}	0.50
{Limes,Asparagus} -> {Large Lemon}	0.54
{Seedless Red Grapes,Limes} -> {Large Lemon}	0.77

Conclusion

- This project helped us learn the application of few data science techniques to the product sales data to
  - (1) predict the product items that might possibly be reordered in future, using standard algorithms in scikit-learn and keras and
  - (2) mine association rules that help discover relation among product items that have the higher probability of being ordered together.
- Challenges in processing big data using ANN and Apriori.