Capstone 2 - Instacart

A Predictive Shopping Experience

Not Your Average Shopping Experience

Instacart is an immersive shopping experience

Users compelled to consume a massive variety of products

Predict which products they will buy again

Data Cleaning & Preparation

Removed one of a kind items

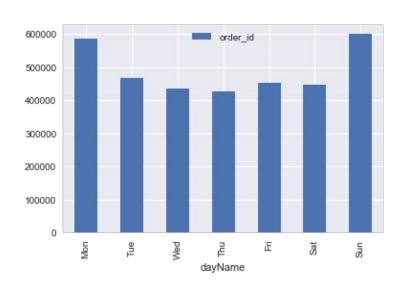
Items with 100% reorder rate or purchased < 40 times in total

Include coding category names for informative plots

Conversion of timestamp data

Aggregative calculations

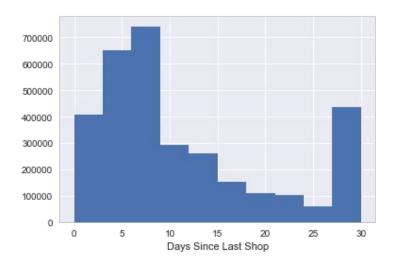
Explorative Analysis

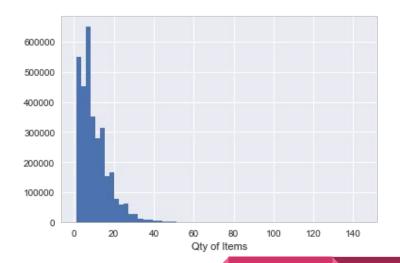




Users are shopping Sunday & Monday, usually midday

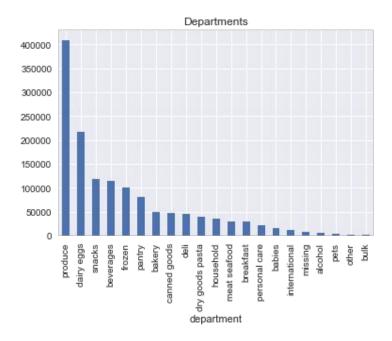
Explorative Analysis





They're shopping every week and buying ~10 items per shop

This is bananas!



Shoppers love produce, dairy, eggs and snacks...

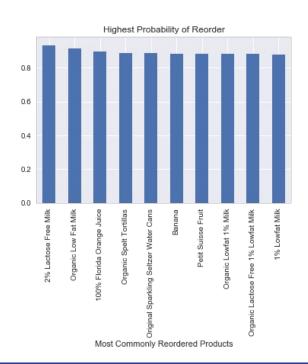


How are these sales reflected by aisle?

produce fresh vegetables		produce packaged vegetables fruits	snacks chips pretzels	snacks	se sp	verages ater Itzer arkling			frozen ice cream ice	frozen frozen produce
			snacks energy granola		be	beverages soft drinks			frozen frozen	
produce fresh fruits			snacks nuts		be	verages			frozen	1 350
		produce	pantry			deli (lunch meat		dry	/ dry	
		produce	pantry							
dairy eggs dairy yogurt milk	dairy eggs	dairy	paintry	De la	Ш	deli				
	milk	eggs eggs	bakery bread	bakery						babies baby
				bakery						baby
dairy eggs packaged cheese	soy	dairy eggs	canned goods							
		dairy eggs	soup							

Explorative Analysis

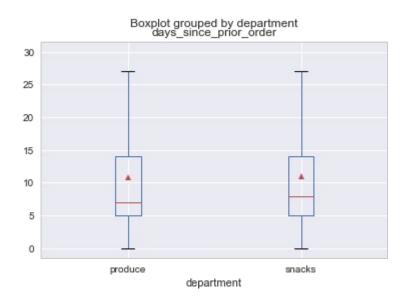
What is most likely to be reordered?





Inferential Stats

Do users reach for snacks before their produce?



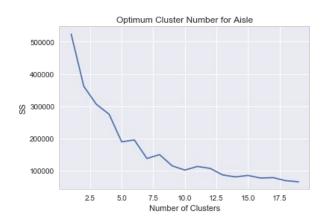
Sample the population

99% Confidence

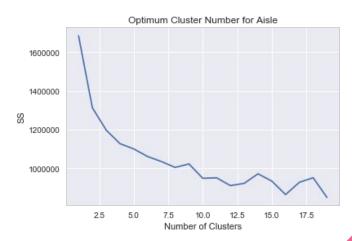
Reject the Null Hypothesis that produce mean days to reorder is equal to snacks

PCA vs Truncated SVD Dimensionality Reduction

Attempted both, explained variance much higher for truncated SVD



PCA Analysis

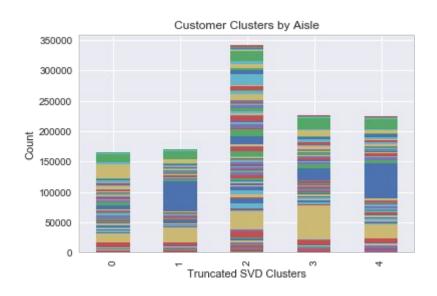


Truncated SVD

Does not centre the data before computing the singular value decomposition

Clustering for customer types

Implemented K-Means on PCA and TruncatedSVD Purchasing Data



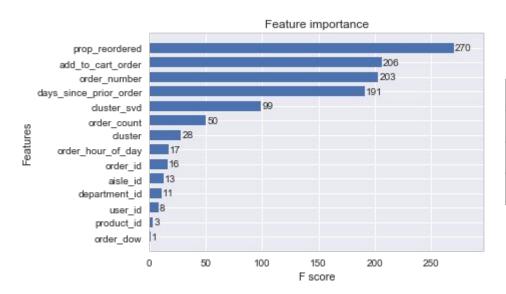
5 Clusters

Similar Product Counts by Aisle

Further Analysis Required...

Modelling & Predictions

XGBoost classifier outperforms them all...



	Logistic	Random Forest	XG Boost		
Precision	0.40	0.67	0.70		
Recall	0.63	0.67	0.71		
F1 Score	0.49	0.67	0.69		

References

Appendix:

Calculations

https://github.com/andrewcmilne/capstone1_instaCart

References

https://www.instacart.com/