Predicting User Engagement

Sonal Bathe

Predicting user engagement to drive retention and sales

- Instacart is an internet-based online grocery delivery service
- It recently released limited user transaction data which can be mined to understand and predict future customer engagement
- The company's strategic goal for the year is to drive customer retention and sales by running a targeted marketing campaign which includes annual membership benefits and free delivery for the subsequent order
- This project would help the company understand and identify its engaged customers to inform the marketing and investment decisions

Business Challenge-

Given a certain marketing campaign budget –

- Which customers should they target?
- In what order?

Machine Learning to the rescue -

Using historic customer transaction information build a model to –

- Predict the “Prime” users
- Rank the customers based on their propensity to be a Prime user

Approach

Data Gathering

- Downloading Instacart data
- AWS EC2 Instance and psql for data storage

Data Preparation

- Data Cleaning
- Feature Engineering
- Defining Target Variable– ‘Prime user’ and assigning binary labels

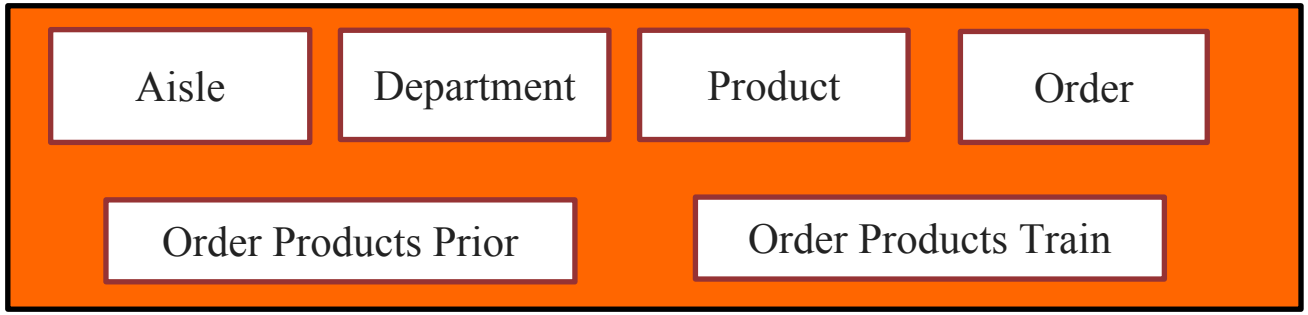
Feature Selection

- Feature selection using a combination of derived features, domain knowledge and the Random Forest Regressor classifier

Modeling

- Logistic Regression
- K Nearest Neighbors
- Gaussian Naïve Bayes
- Decision Tree
- Support Vector Machine

Instacart Data -



Join in psql

user_id	order_id	department_	department	aisle_id	aisle	product_id	product_nar	eval_set	order_num	order_dow	order_hour_	days_since_prior	add_to_cart_order	reordered
1	1187899	19	snacks	121	yogurt	49235	Zero Calorie	train	11	4	8	14	11	1
2	1492625	20	snacks	117	soy lacto	48821	Vegetable La	train	15	1	11	30	31	1
5	2196797	17	produce	123	packagec	48204	Tamari Glute	train	5	0	11	6	9	1

Feature Engineering – Derived Variables

- Average orders per user
- Average items per order per user
- Average reordered items per order per user
- Average days since prior order per user
- Max days since prior order per user
- Day of the week user orders more frequently
- Hour of the day user orders more frequently

Feature Selection

22 Input Features

Determined based on:

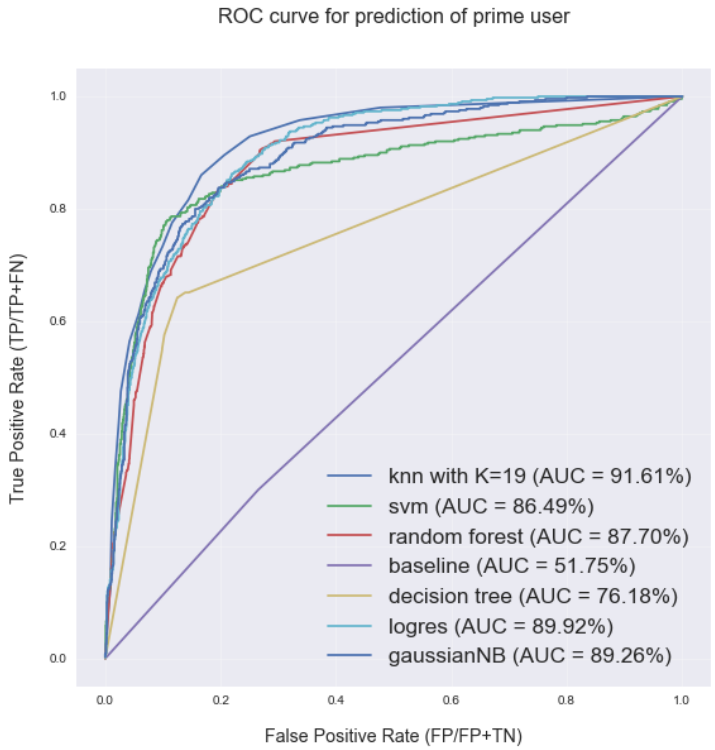
- Correlation between variables
- Referring to Random Forest Regressor generated important features

4 Selected Features :

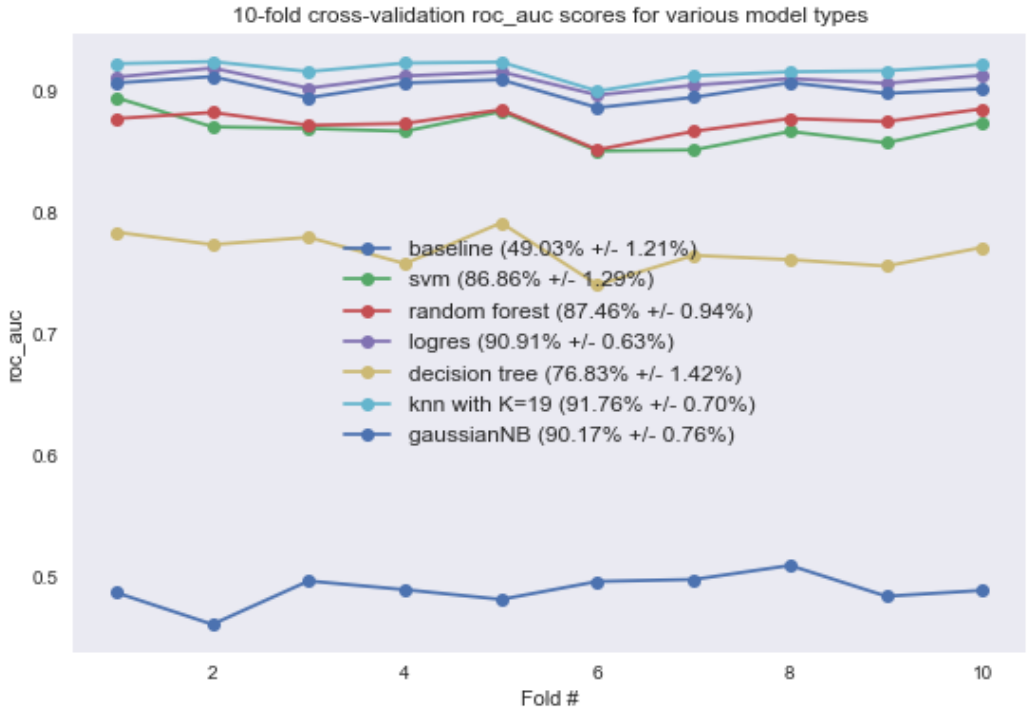
- Average reordered items per order per user
- Average days since prior order per user
- Day of the week user orders more frequently
- Hour of the day user orders more frequently

➢ **Created a target variable – 'Prime User'.** Top 25% of the users based on the most predictive features were assigned a label of 1

Comparison of models using cross validation

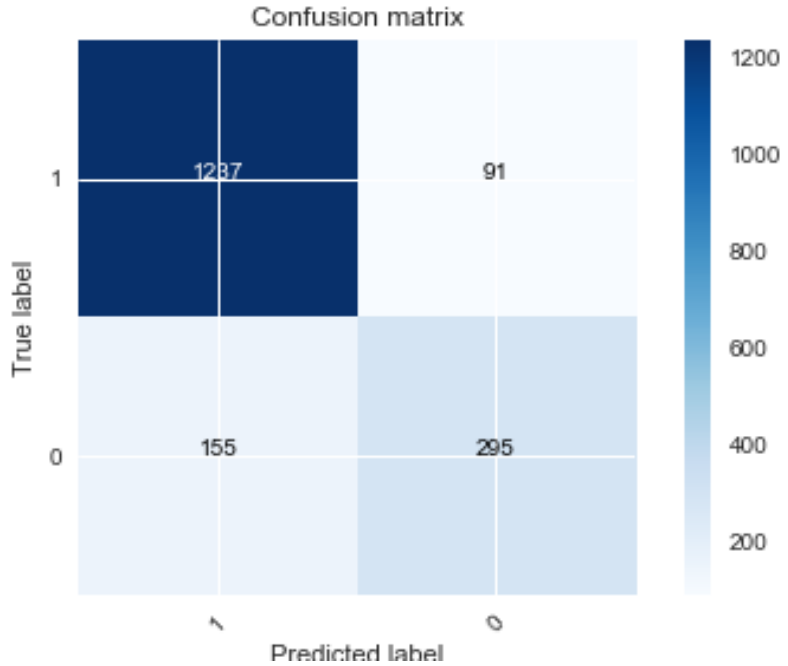


Model comparison by folds



Performance metrics for the best model

Metric	Score
Accuracy :	0.861
Recall :	0.861
Precision :	0.857
ROC_AUC :	0.916



Recommendation

- To enable an informed decision – Use the best model to rank the customers and select those with higher probability of being a Prime user
- Business to decide on model threshold based on their appetite for precision and recall trade-off
- Depending on the budget of the campaign it is advisable to conduct a test/control setup for the selected customer base to evaluate campaign effectiveness

Next Steps

- As and when Instacart releases the pricing data for the products as well as the data for the dollar value brought in by each user, the model predictions can be further improved.

- The company revenue can be maximized by considering the average cost for each targeted prime user and the average revenue per targeted prime user

THANKS!

Any questions?