STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Predicting Customer Loyalty

Final Project: Marketing Analytics

By:
Vaibhav Prasad Desai
Abhinav S Panwar
Piyush Bhattad
Nishad Gawde

AGENDA

- Introduction
- Dataset Details
- Feature Engineering with Google Big Query
- Machine Learning with Microsoft Azure
- Results
- Conclusion

OBJECTIVE

- Let's say 50 customers are offered a discount to purchase two bottles of soda. Out of the 50 customers, 10 choose to redeem the offer. These 10 customers are the focus of this project.
- The Objective is to predict whether the customer who has been given offer will redeem the offer in the near future.
- The training dataset has a binary field 'Repeater', which is response variable.
- To create this prediction, it was given a minimum of one year of shopping history prior to each customer's incentive, as well as the purchase histories of many other shoppers

INTRODUCTION

- Companies often offer discounts to attract new shoppers to buy their products. The most valuable customers are those who return after this initial purchase
- With enough purchase history, it is possible to predict which shoppers, when presented an offer will buy a new item
- We have used **Acquire Valued Shoppers Data** from Kaggle .It contains almost 350 million rows of completely anonymized transaction data from over 300,000 shoppers.
- The dataset was loaded on Google Cloud as the dataset was very huge in size(22 GBs).
- Big Query and Azure Machine Learning platform were used for statistical analysis in this project

DATA

These are the four relational files in the dataset

- transaction.csv- contains transaction history for all customers for a period of at least a year prior to their offered incentive
- trainHistory.csv - contains the incentive offered to each customer and information about the behavioral response to the offer
- testHistory.csv - contains the incentive offered to each customer but does not include their response
- offers.csv - contains information about offer

DATA

All of the fields are anonymized and categorized to protect customer and sales information

HISTORY	
id	A unique is representing a customer
Chain	An integer representing a store chain
Offer	An id representing a certain offer
Market	An id representing a geographical region
Repeattrips	The number of times the customer made a repeat purchase
Repeater	A Boolean, equal to repeattrips > 0
Offerdate	The date a customer received the offer

TRANSACTIONS	
Id	A unique is representing a customer
Chain	An integer representing a store chain
Dept.	An aggregated grouping of the category
Category	The product category
Company	An id of the company that sells the item
Brand	An id of the brand to which the item belongs
Date	The date of purchase
Productsize	The amount of the product purchase
Productmeasure	The units of the product purchase
Purchasequantity	The number of units purchased
Purchaseamount	The dollar amount of the purchase

DATA

OFFERS	
Offer	An id representing a certain offer
Category	The product category
Quantity	The number of units one must purchase to get the discount
Company	An id of the company that sells the item
Offervalue	The dollar value of the offer
Brand	An id of the brand to which the item belongs

FEATURE ENGINEERING

- Many companies have collected and stored huge amount of data about their current, past and potential customers, suppliers and business partners.
- However, the inability to discover valuable information hidden in the data prevents the companies from transforming these data into valuable and useful knowledge.
- Data mining tools could help these companies to discover the hidden knowledge in the enormous amount of data
- “Company”, “Category” and “Brand” represent the most important characteristics in the transaction data.

FEATURE ENGINEERING

Secondary features:

id
repeattrips
repeater
offerdate
Recency
monetary
Frequency
Bght_Comp_or_Not
has_bght_category_or_not
has_bght_brand_or_not
has_bght_comp_cat_brand_together_not
has_bght_comp_brand_together_not
has_bght_comp_cat_together_not
has_bght_cat_brand_together_not
offervalue

No_of_trips_Company
Tot_Purch_comp
Tot_amt_comp
No_of_trips_company_30
Tot_Purch_Company_30
Tot_amt_comp_30
No_of_trips_company_60
Tot_Purch_Company_60
Tot_amt_comp_60
No_of_trips_company_90
Tot_Purch_Company_90
Tot_amt_comp_90
No_of_trips_company_180
Tot_Purch_Company_180
Tot_amt_comp_180

FEATURE ENGINEERING

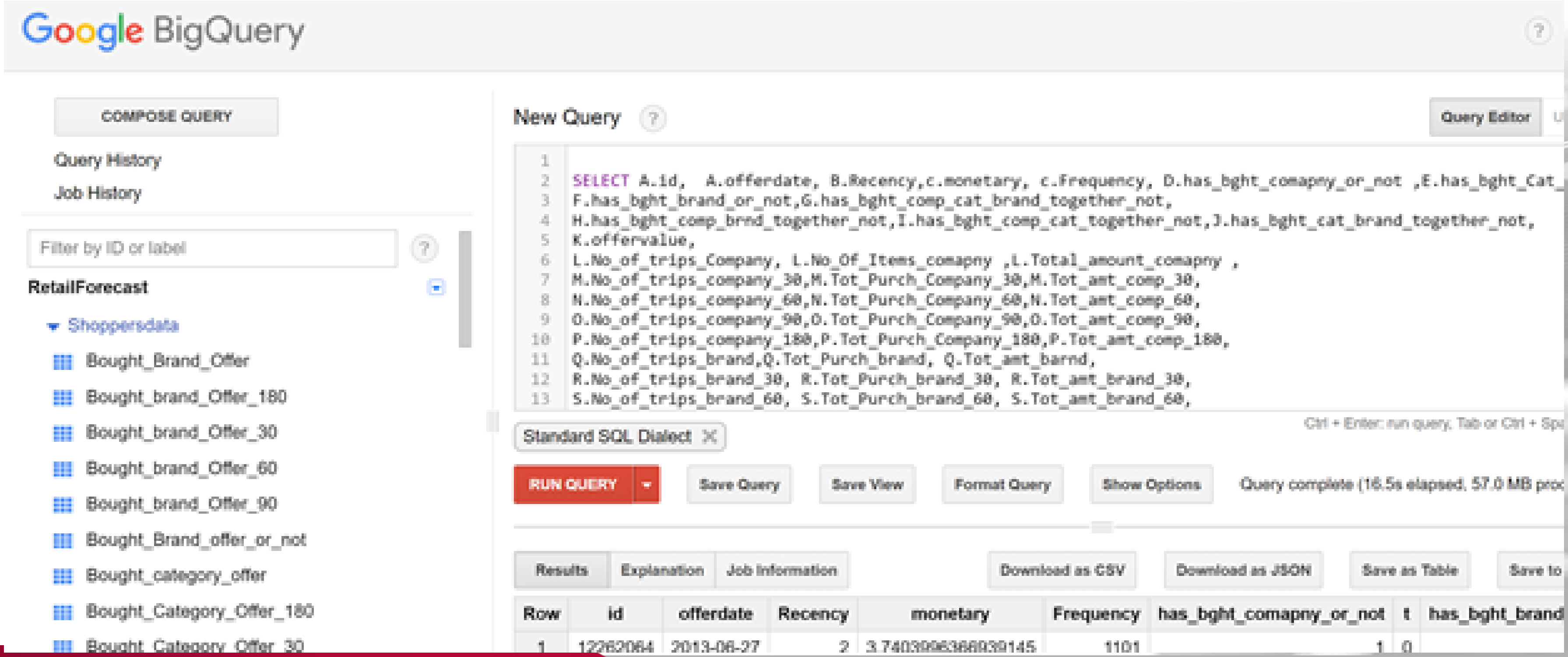
Secondary features:

No_of_trips_brand
Tot_Purch_brand
Tot_amt_brand
No_of_trips_brand_30
Tot_Purch_brand_30
Tot_amt_brand_30
No_of_trips_brand_60
Tot_Purch_brand_60
Tot_amt_brand_60
No_of_trips_brand_90
Tot_Purch_brand_90
Tot_amt_brand_90
No_of_trips_brand_180
Tot_Purch_brand_180
Tot_amt_brand_180

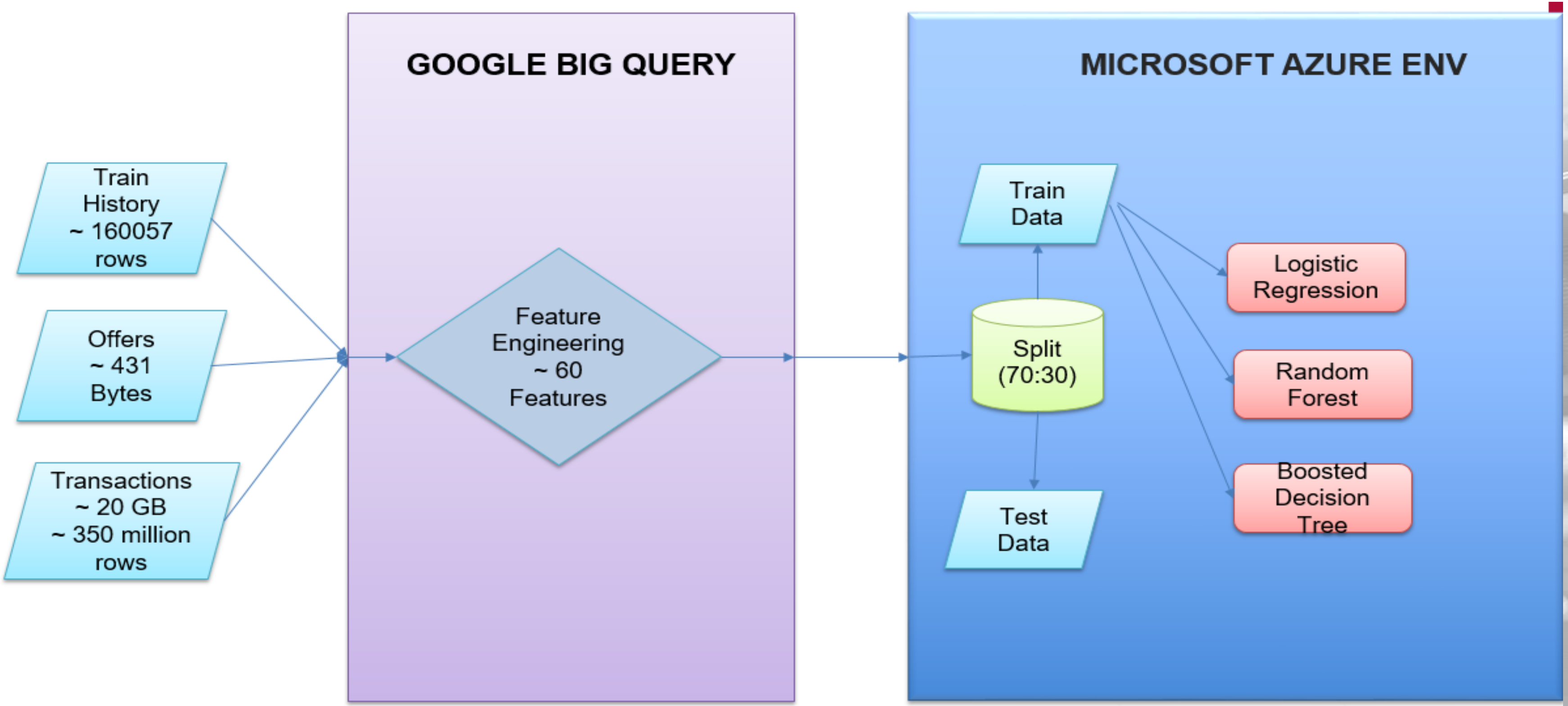
No_of_trips_Category
Tot_Purch_cat
Tot_amt_cat
No_of_trips_cat_30
Tot_Purch_Cat_30
Tot_amt_cat_30
No_of_trips_cat_60
Tot_Purch_Cat_60
Tot_amt_cat_60
No_of_trips_cat_90
Tot_Purch_Cat_90
Tot_amt_cat_90
No_of_trips_cat_180
Tot_Purch_Cat_180
Tot_amt_cat_180

FEATURE ENGINEERING

Big Query Interface:



METHODOLOGY



Google Big Query(SQL)

CSV Files(structured Data) Imported as Tables

Using SQL and power of Big Query to handle
The big data(350 million rows) for
feature engineering.

```
select A.id ID , count(A.ID) No_of_trips_company_30,
Sum(A.purchasequantity) Tot_Purch_Company_30
,Sum(A.purchaseamount) Tot_amt_comp_30
from [retailforecast:Shoppersdata.transactions] A
inner join [retailforecast:Shoppersdata.Trainhistory] B
 on A.id=B.id inner join  [retailforecast:Shoppersdata.offers] C
 ON B.offer=C.offer and A.company=C.company
 where DATE(A.date)> DATE(DATE_ADD(B.offerdate, -30, 'DAY') )
Group by ID
Order by ID
```

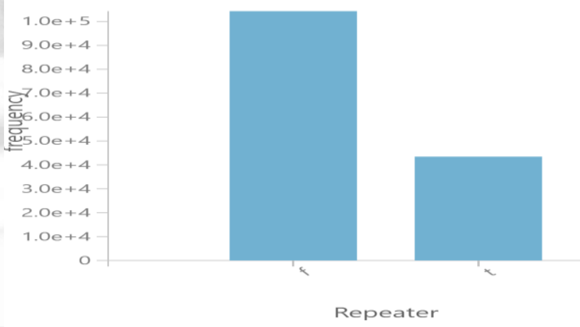
Data Processing

Import Data -> Azure Cloud dataset

<u>Exploratory Data Analysis:</u>
Variable Type Identification,
Univariate Statistics
Histogram Plots

<u>Data Cleaning:</u>
Casting columns/Project Columns
NA Values =>0
Response variable mapped to Binary 1&0

Feature	Count	Unique Value Count	Missing Value Count	Min	Max
Total_sum	160057	159525	0	-5953.67	48321663.58999
Total_count	160057	4314	0	1	2647164
Total_quant	160057	6716	0	-28429	18742274
Tot_comp_sum	160057	17101	0	-28.47	90477.57
Tot_comp_count	160057	163	0	0	9562
Tot_comp_quant	160057	284	0	-3	29585



Azure Machine Learning

Model Development:

Random Split: Train (0.7) & Test (0.3)

Response (binary)~Primary features+ Secondary Features

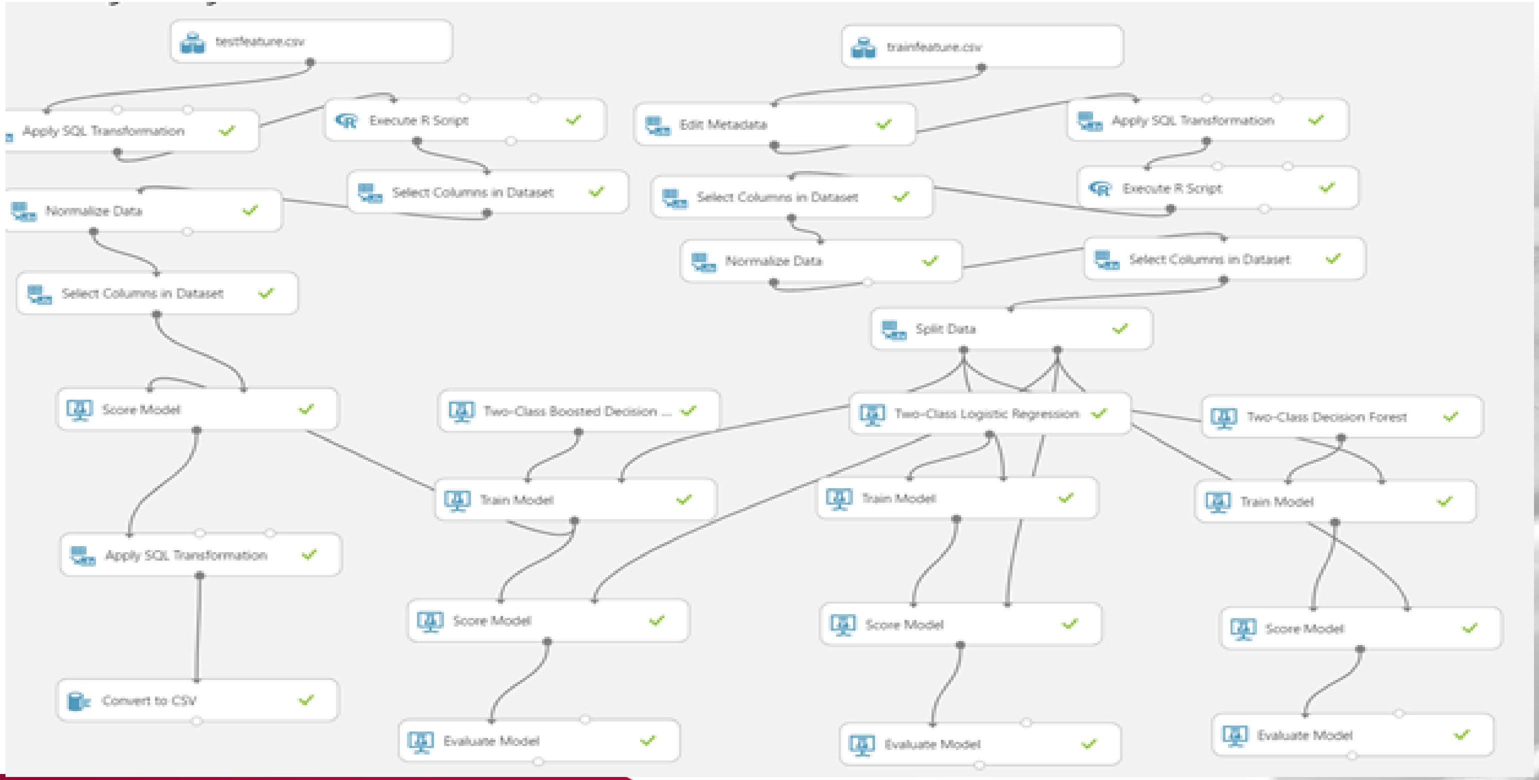
Label Classification ML Algorithms:

- Logistic Regression
- Boosted Decision Tree

Relationships with dependent variable(influence)

repeater	has_bought_cat	offer_maade_order	has_bought_brand	Tot_cat_count
1	0.127751	0.122635	0.11307	0.066667

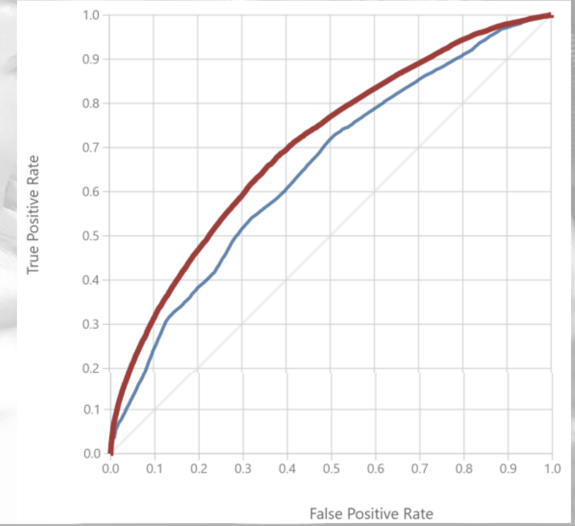
Azure Machine Learning



Results

ML Model	True +ve	False +ve	True -ve	False -ve	Accuracy	Precision	AUC
Logistic	394	118	34876	12629	0.735	0.77	0.631
Boosted decision Forest	2997	2104	32890	10026	0.747	0.588	0.704

Receiver Operating Curve =>



Learnings

- Feature engineering plays an important role in improvement of prediction
- Cloud services like Google Big Query and Microsoft Azure ML works efficiently with big datasets.
- Azure ML also gives a platform for data scientists to automate the tedious work of cleaning and data processing along with power of running, tuning,training and, testing different models simultaneously.

Thereby the best model after several training & tuning can be used to design offers for customers more efficiently and further can also be used for generating Lifetime Value.

Thank You!