E-COMMERCE ANALYSIS WITH GOOGLE MARKETPLACE

JUSTINE PICAR, OCTOBER 2021

BUSINESSES AND E-COMMERCE: PROBLEM

- Businesses fail due to:
  - Lack of research
  - Not the right market
  - Not reaching the right people
- 21.5% fail within the 1st year
- 30% fail within the 2nd year
- 50% fail within the 5th year
- 70% fail within the 10th year
- For online businesses, 90% fail within 4 months

PROJECT OBJECTIVE:

Identify what metrics will improve profits for online stores by exploring and analyzing user propensity to purchase and provide recommendations for maintaining customer loyalty and mitigating attribution issues

DATA: GOOGLE MERCHANDISE DATA FROM BIGQUERY REST API

- INCLUDES TRAFFIC SOURCE, CONTENT, TRANSACTIONAL DATA
- CONSTRAINTS: HIDDEN, REMOVED, AND/OR DEPRECATED FIELDS
- DATA CAN ONLY BE QUERIED OR USED TO GENERATE REPORTS

DATA WRANGLING

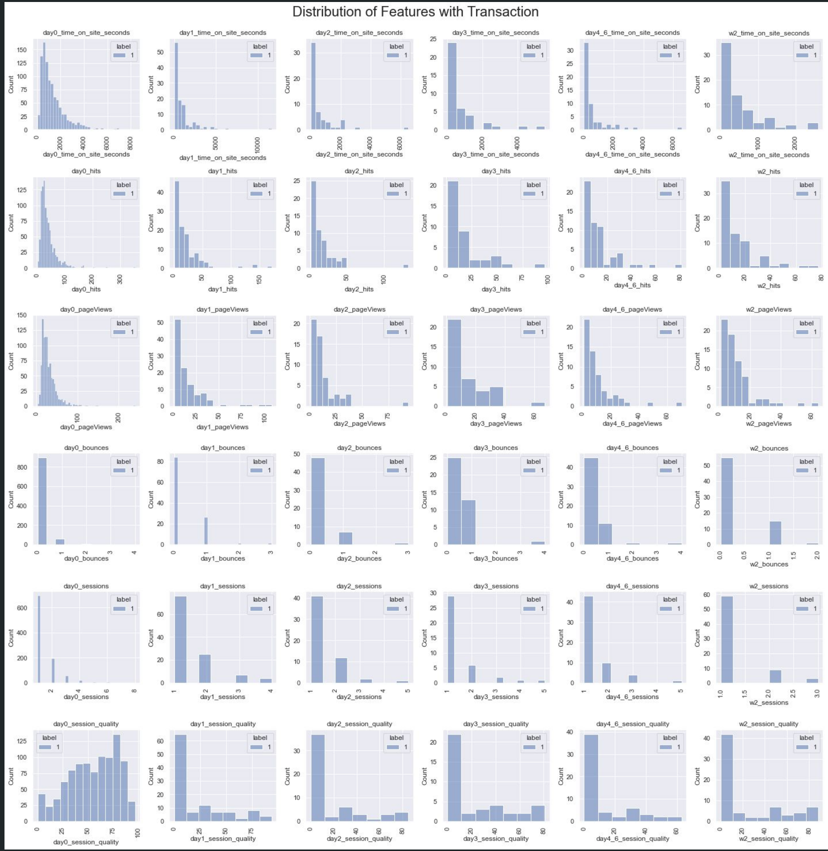
- Used Standard SQL to query the data
- Identified useful features:
  - *fullVisitorId*
  - *date*
  - *visits or sessions*
  - *hits*
  - *pageviews*
  - *bounces*
  - *sessionQuality*
  - *timeOnSite* (in seconds)
- Unnested *totals* column to aggregate the features by day over the course of two weeks and create new features
- Binary Target Variable (Transactions)

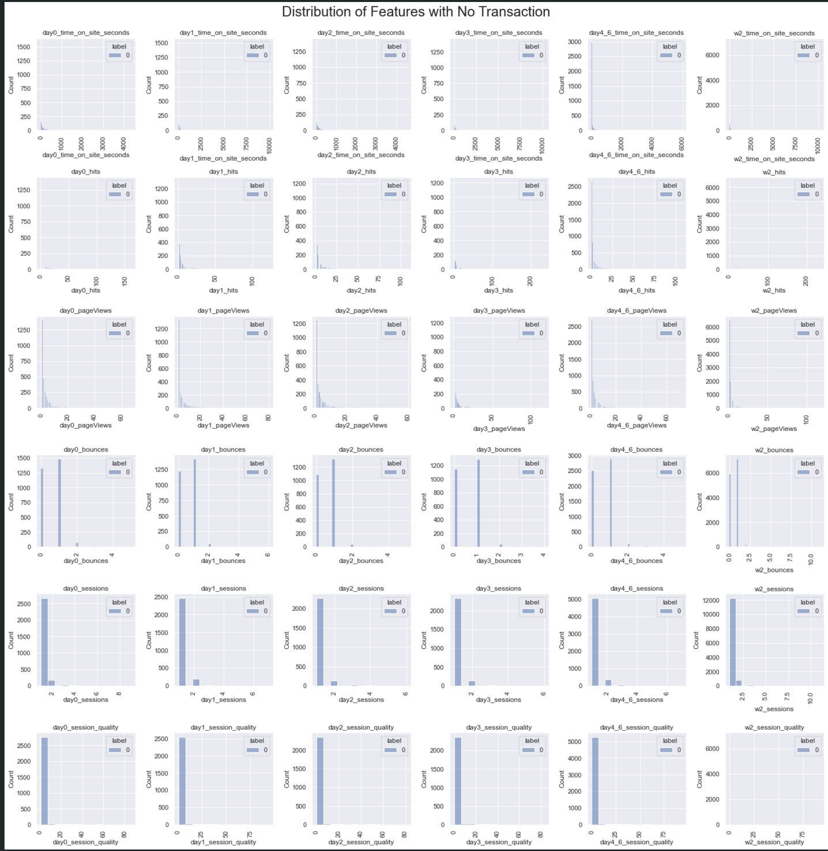
- Features broke out in the following format from 12am to 11:59pm:
  - Day 0 - day of purchase
  - Day 1 - 1 day before purchase
  - Day 2 - 2 days before purchase
  - Day 3 - 3 days before purchase
  - Day 4-6 - 4-6 days before purchase
  - Week 2 - 2 weeks before purchase
- Total of ~20k values for train set, ~19k validation set with 38 total columns
- Training Set from 07/1/2017 - 07/31/2017
- Test Set from 03/1/2017 - 03/14/2017
- Each feature is grouped by a distinct and unique visitor id over 10 days

PACKAGES

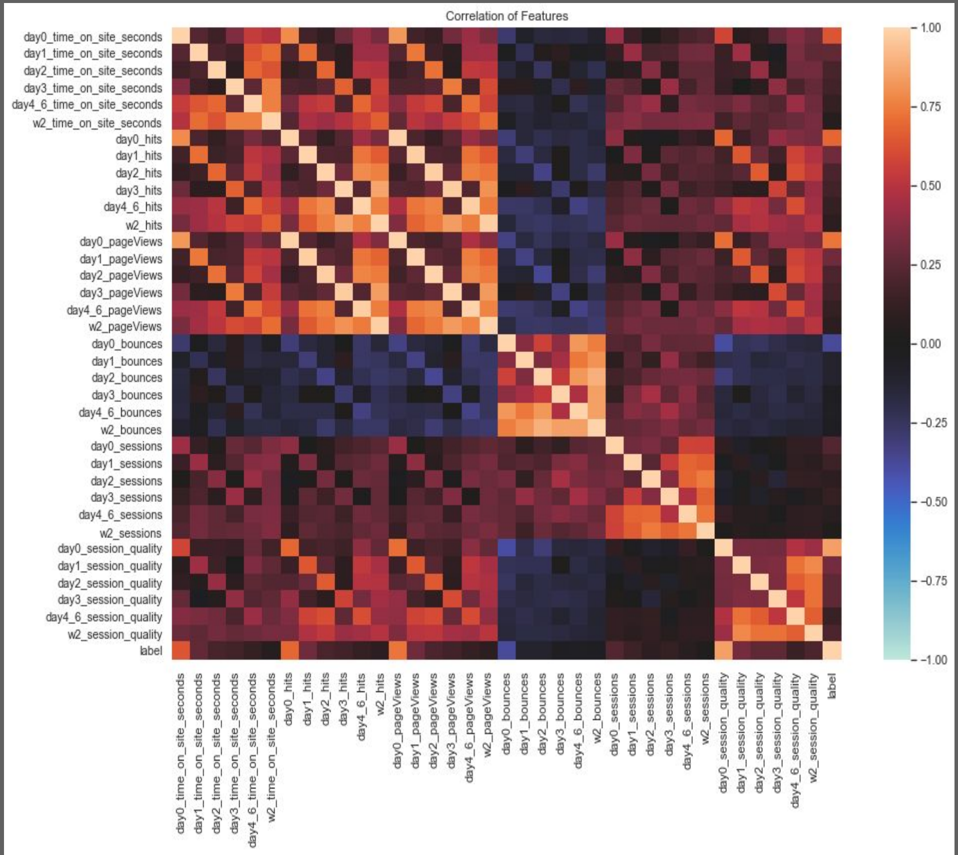
- For preprocessing, modeling the data, and calculating metrics
  - sklearn
- For data and metric visualization
  - matplotlib.pyplot
  - skplot
  - seaborn
  - shap
- For querying the data from API to jupyter notebook
  - google.cloud

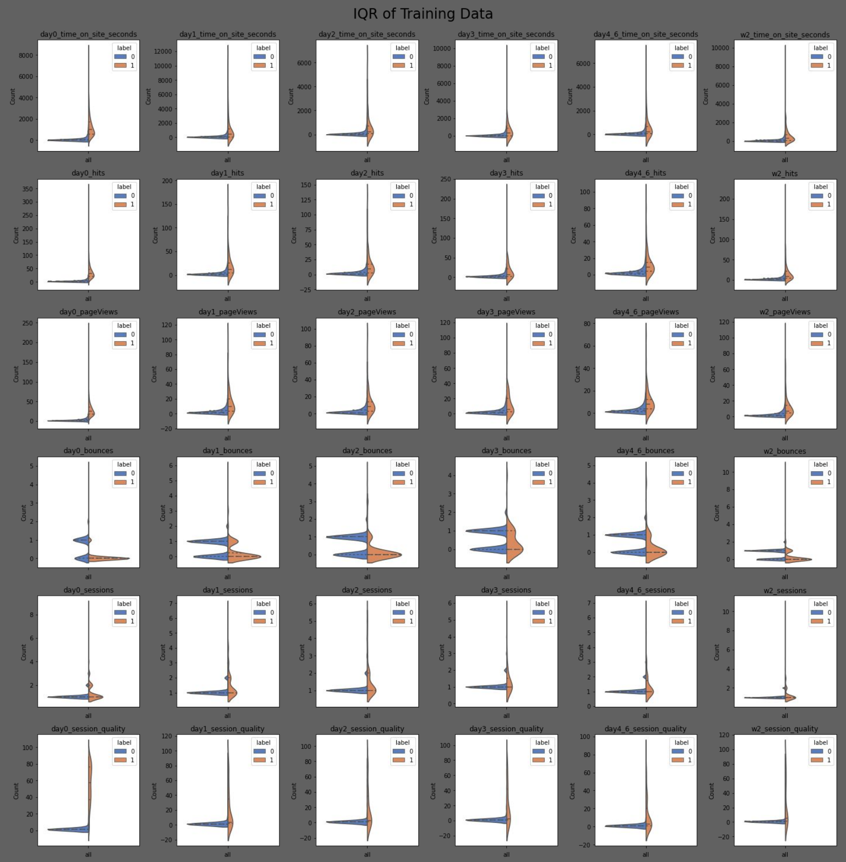
EXPLORATORY DATA ANALYSIS: DISTRIBUTIONS





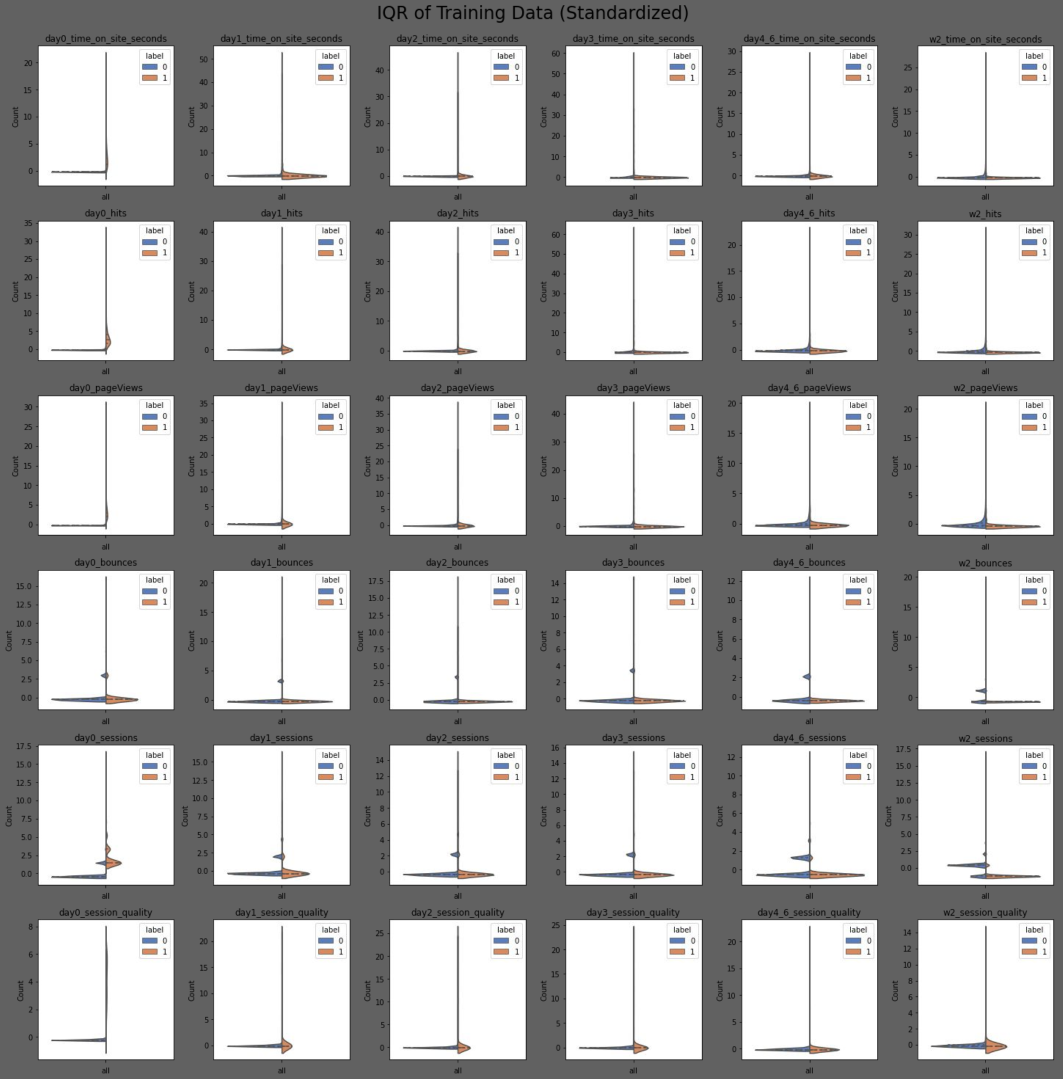
EXPLORATORY DATA ANALYSIS: FEATURE IMPORTANCE AND IQR



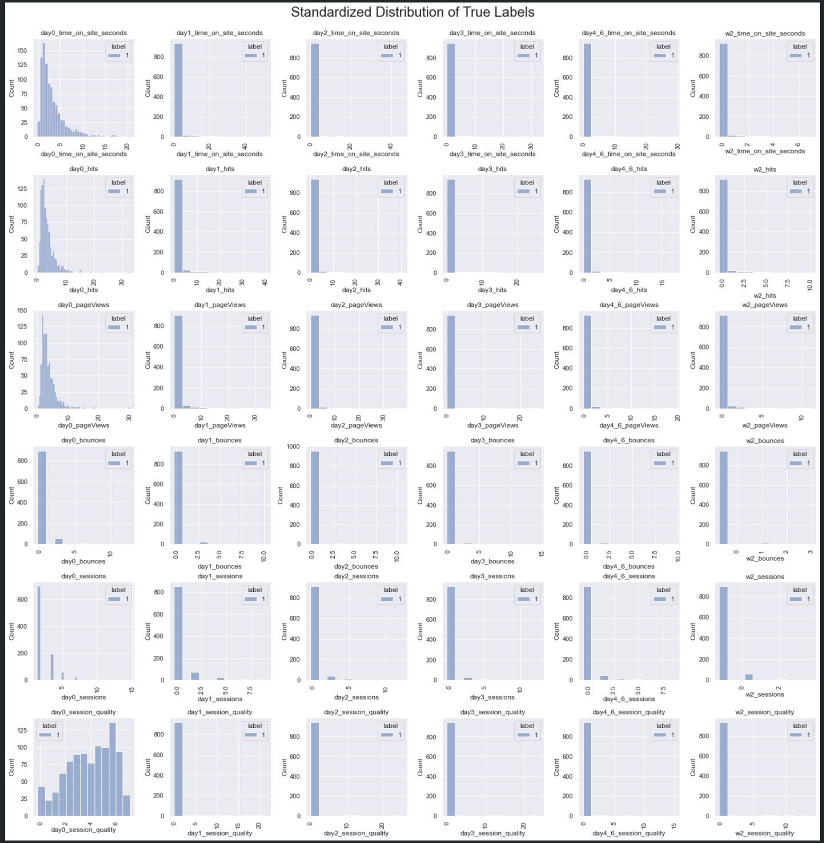


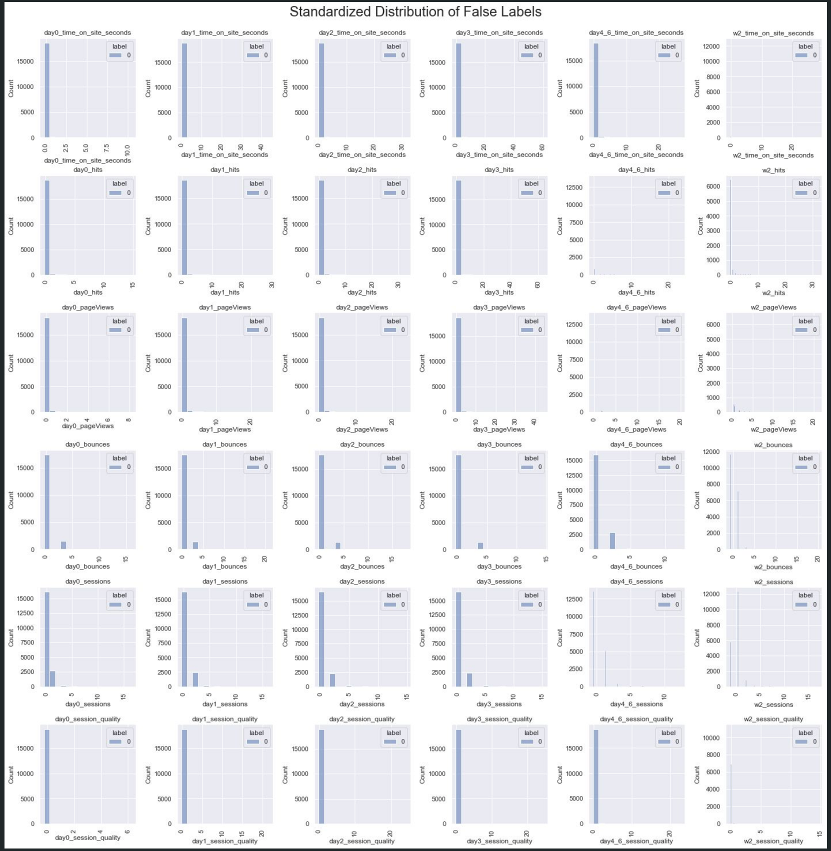
PRE-PROCESSING: SCALING AND IMPUTING THE DATA

- Increase robustness of models and ability to generalize data
- Only 5% of data are true transactions; highly skewed
- Normalize data with Standardization
- Impute null values with zero (this represents time spent on site when visitor was not present)



PRE-PROCESSING: SCALING DISTRIBUTION OF DATA

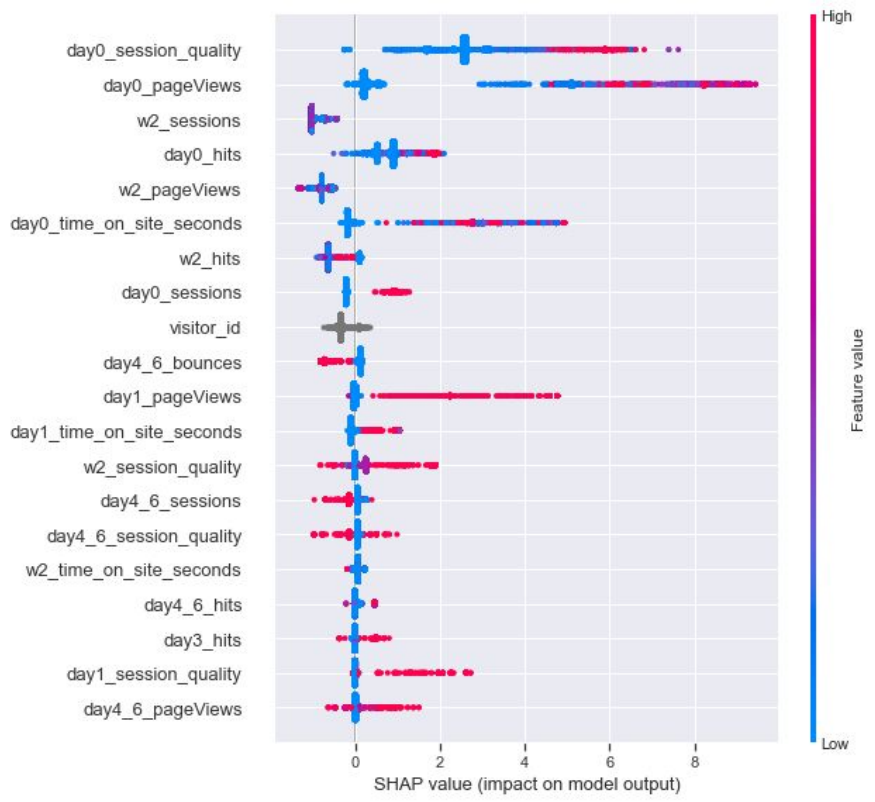


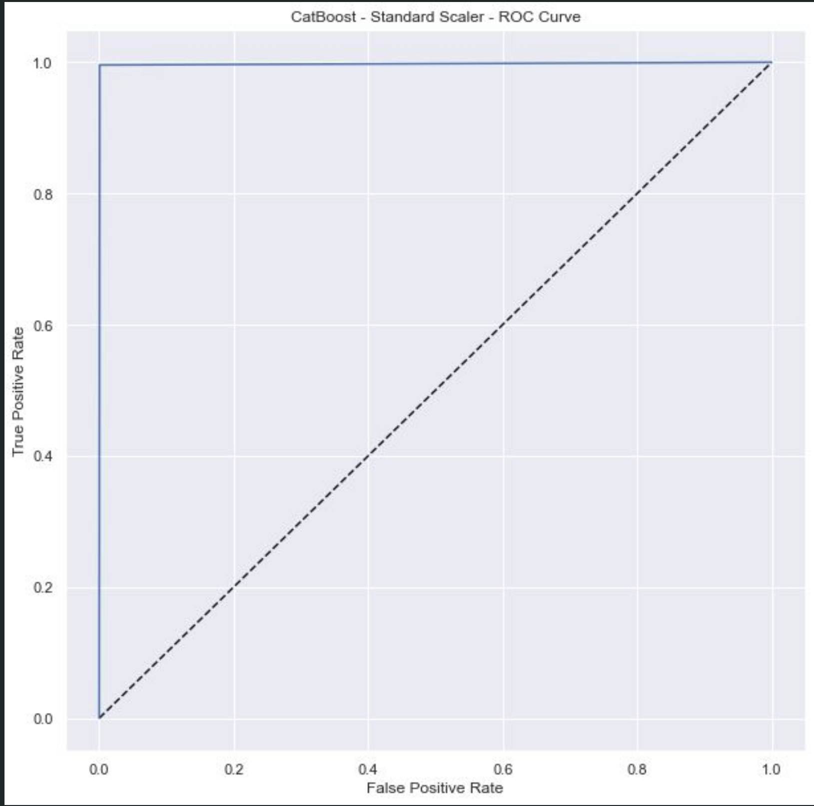


Methodology

- Depending on the goals of the company, certain features included may be dropped or added as needed
- The model in its current form is simplified
- Has no consideration for platform constraints which would need to be defined by the company
- Models used:
    - CatBoost
    - Logistic Regression
    - Random Trees
- Metrics considered
    - SHAP value (impact on model output) for CatBoost or feature importance on Random Forest and Logistic Regression
    - True Positive Rate from the ROC curve
    - LIFT score for customer segmentation

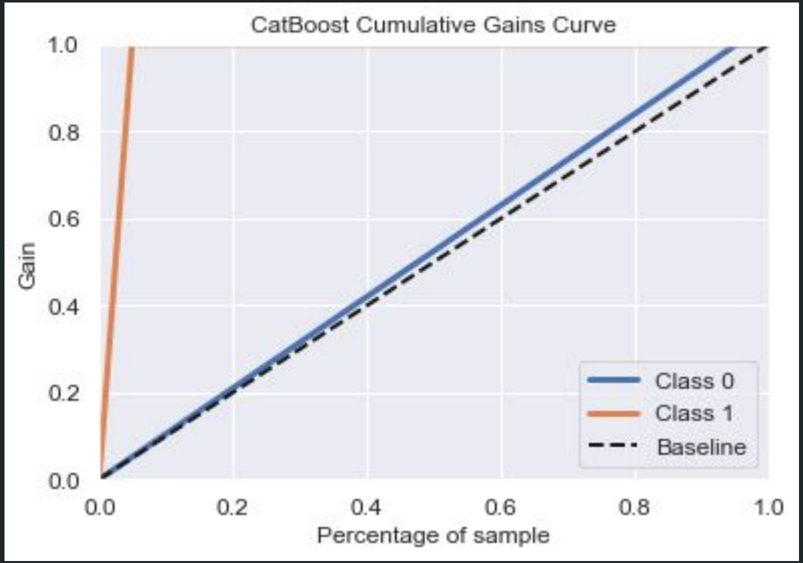
MODELING: CATBOOST, ACCURACY: 99.91%



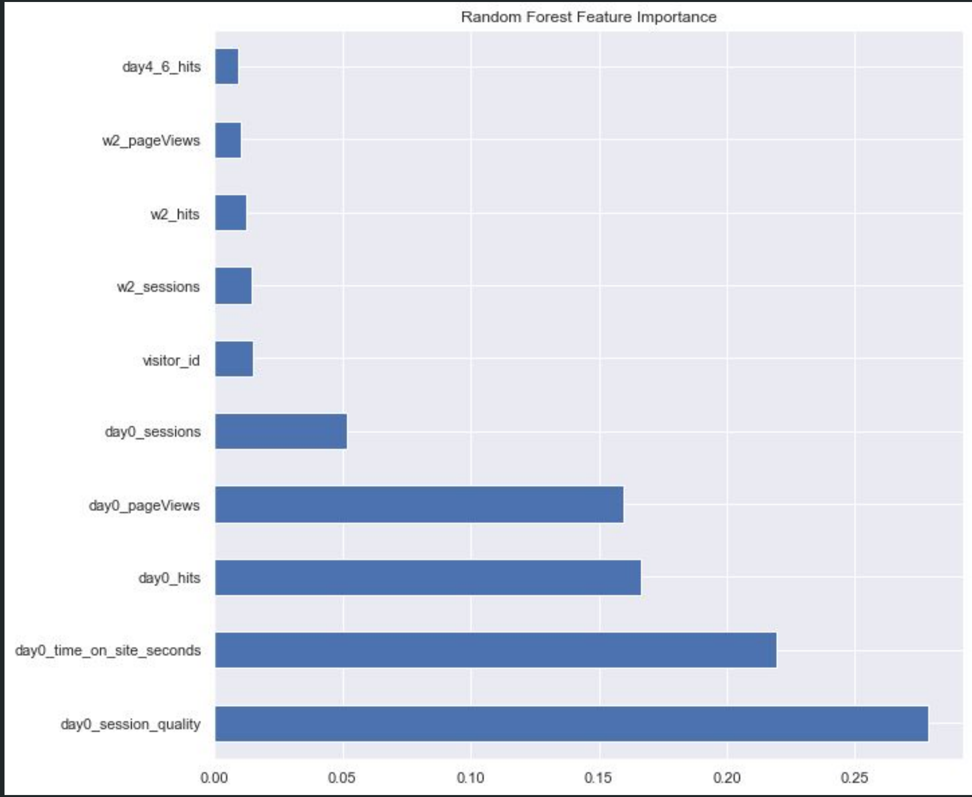


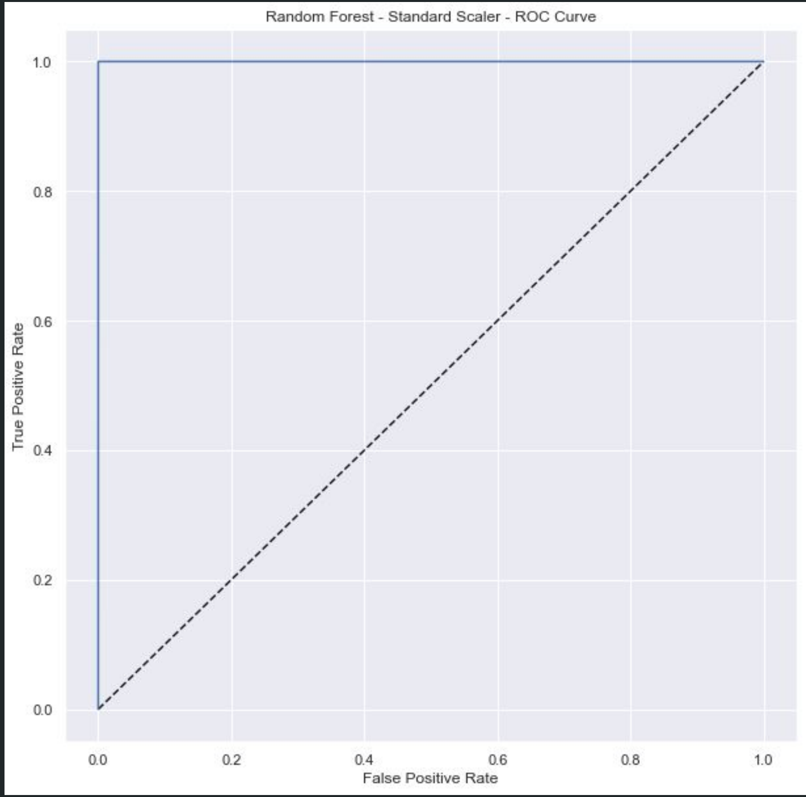
MODELING: CATBOOST, ACCURACY: 99.91%

	label_neg	ntile	count_neg	label_pos	count_pos	lift
0	0	1	1997	1	0	0.000000
1	0	2	1996	1	0	0.000000
2	0	3	1996	1	0	0.000000
3	0	4	1997	1	0	0.000000
4	0	5	1996	1	0	0.000000
5	0	6	1996	1	0	0.000000
6	0	7	1997	1	0	0.000000
7	0	8	1996	1	0	0.000000
8	0	9	1996	1	0	0.000000
9	0	10	1024	1	973	18.545898



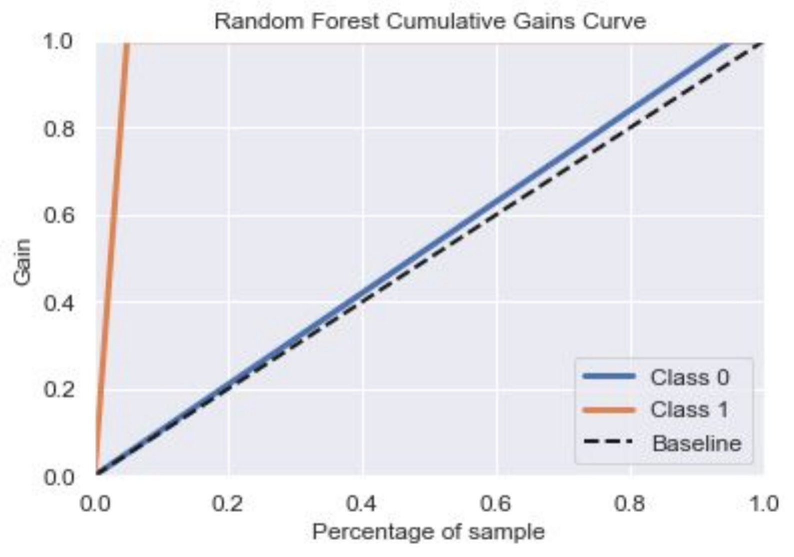
MODELING: RANDOM FOREST, ACCURACY: 99.99%



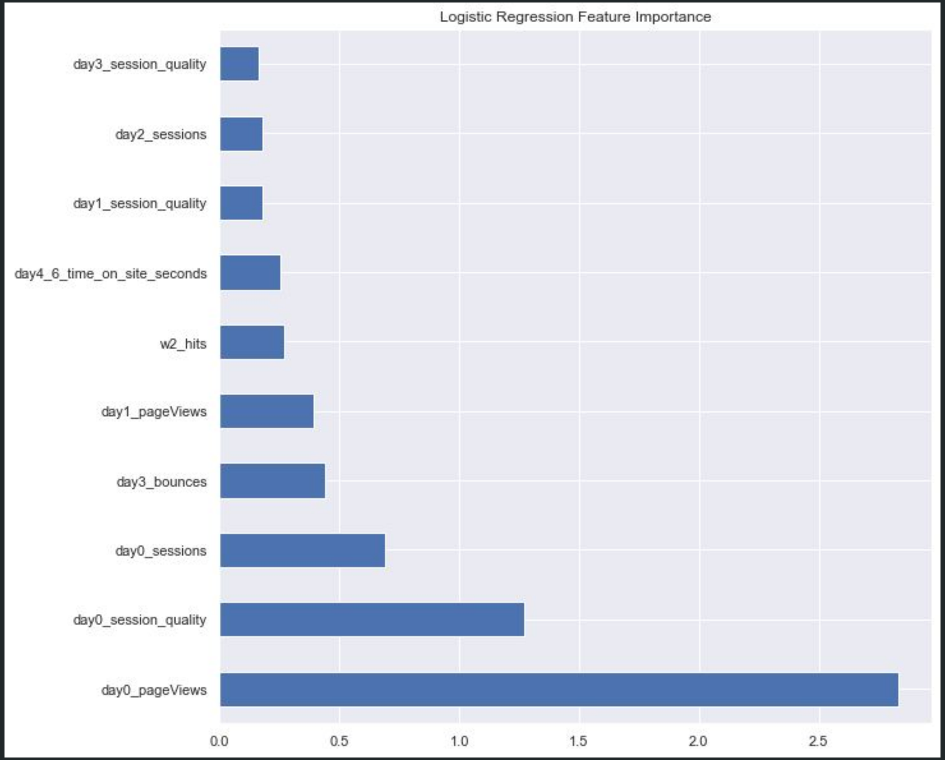


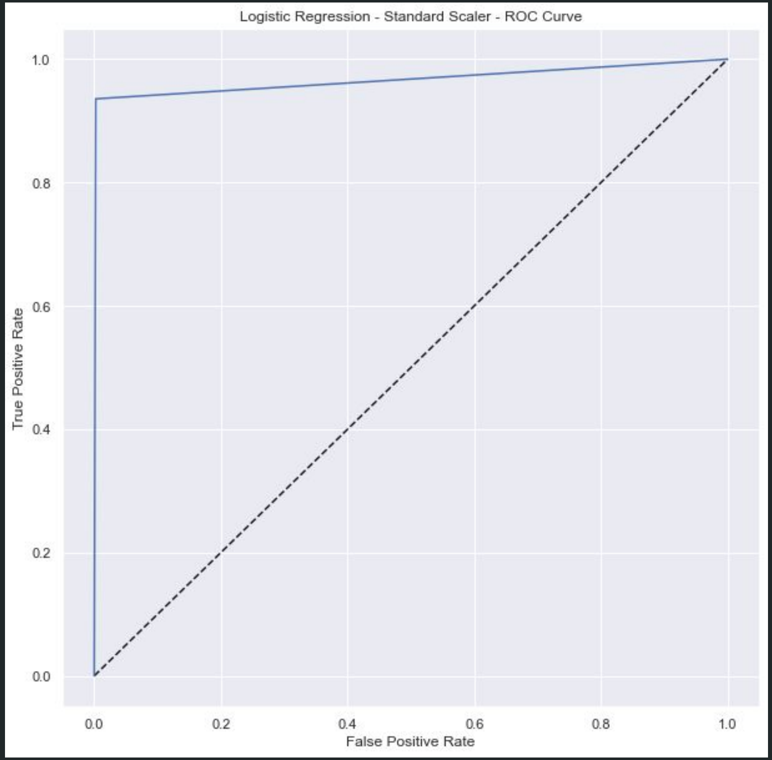
MODELING: RANDOM FOREST, ACCURACY: 99.99%

	label_neg	ntile	count_neg	label_pos	count_pos	lift
0	0	1	1997	1	0	0.000000
1	0	2	1996	1	0	0.000000
2	0	3	1996	1	0	0.000000
3	0	4	1997	1	0	0.000000
4	0	5	1996	1	0	0.000000
5	0	6	1996	1	0	0.000000
6	0	7	1997	1	0	0.000000
7	0	8	1996	1	0	0.000000
8	0	9	1996	1	0	0.000000
9	0	10	1032	1	965	18.409884



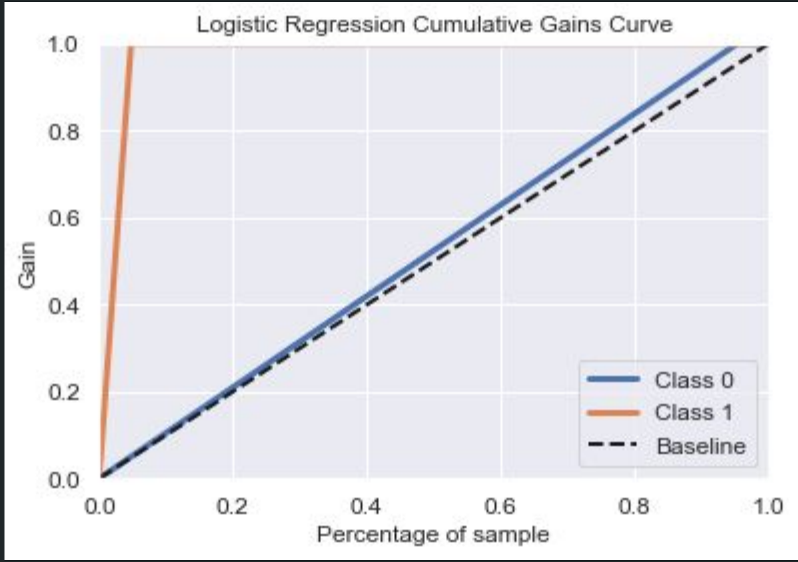
MODELING: LOGISTIC REGRESSION, ACCURACY: 100%





MODELING: LOGISTIC REGRESSION, ACCURACY: 100%

	label_neg	ntile	count_neg	label_pos	count_pos	lift
0	0	1	1997	1	0	0.000000
1	0	2	1996	1	0	0.000000
2	0	3	1996	1	0	0.000000
3	0	4	1997	1	0	0.000000
4	0	5	1996	1	0	0.000000
5	0	6	1996	1	0	0.000000
6	0	7	1997	1	0	0.000000
7	0	8	1996	1	0	0.000000
8	0	9	1996	1	0	0.000000
9	0	10	1050	1	947	18.111429



RESULTS AND ANALYSIS

- BEST MODEL: CATBOOST
  - Highest lift score, successfully identified most users who purchase
  - True Positive Rating similar to Random Forest Model
    - Random Forest lift function is lower
- WORST MODEL: LOGISTIC REGRESSION
  - 100% accuracy, but lowest lift score and lowest true positive rating
  - Suggests there may be overfitting occuring
- Feature Importance:
  - Users engagement increases as they approach the purchase date (day0 features); transactions and activities are highly correlated to the day of purchase
    - Page views, session quality, and time on site have the highest positive correlations
    - Bounces have a negative correlation

APPLICATIONS

- Identify segments of users likely to organically convert
  - Allocate budget to market to this segment and mitigate attribution
  - Optimize email and marketing campaigns to specific users
- Provide discounts for other segments less likely to convert
  - Encourage users to make purchases
- Paid media marketing campaigns (on Amazon, Facebook, Instagram, etc.)
- Customer Service: Reroute and prioritize calls based on user's propensity to purchase

CONCLUSION & LESSONS LEARNED

- 99.91% ACCURACY ON CATBOOST MODEL (note: for purpose of justifying roc, lift, other metrics, etc)
- PRE-PROCESSING
  - Upsampling smaller labels
  - Imputing mean or median
  - Normalize using min-max scale or log transformation
- MODELING:
  - ADA BOOST
  - XG BOOST
- SELECT MORE FEATURES FROM BIGQUERY TO EXPLORE

THANK YOU!

Thank you Nik for being my mentor! It’s been wonderful working with you.

SOURCES

- https://www.investopedia.com/articles/personal-finance/040915/how-many-startups-fail-and-why.asp
- https://www.lendingtree.com/business/small/failure-rate/
- https://www.huffpost.com/entry/10-reasons-why-your-new-online-business-will-fail_b_7053610
- https://console.cloud.google.com/marketplace/product/obfuscated-ga360-data/obfuscated-ga360-data?project=lexical-script-761
- https://support.google.com/analytics/answer/3437719?hl=en
- Stack Overflow
- https://www.analyticsvidhya.com/blog/2017/08/catboost-automated-categorical-data/