LEVI’S SENIOR DATA SCIENTIST PROJECT - EMAIL EFFECTIVENESS

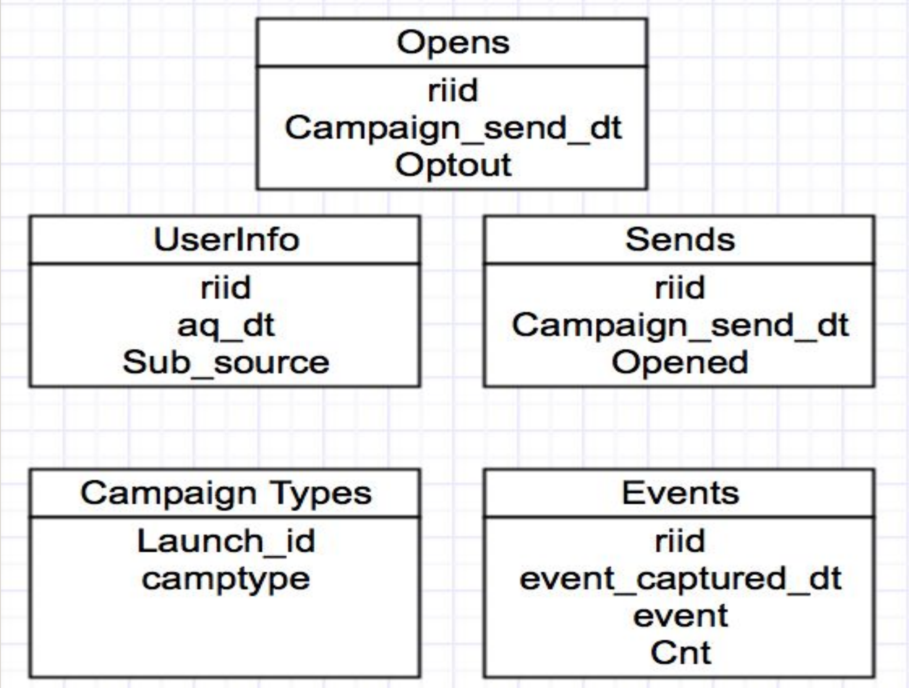
SAI HARSHITH REDDY GADDAM

Problem statement

Increase the effectiveness of email sends by identifying the right set of target customers for the upcoming campaigns

- Predict the likelihood that a customer will open an email sent on a given day in the future for a given campaign type
- Predict the likelihood that a customer that opens will unsubscribe

Data provided



Opens
riid
Campaign_send_dt
Optout

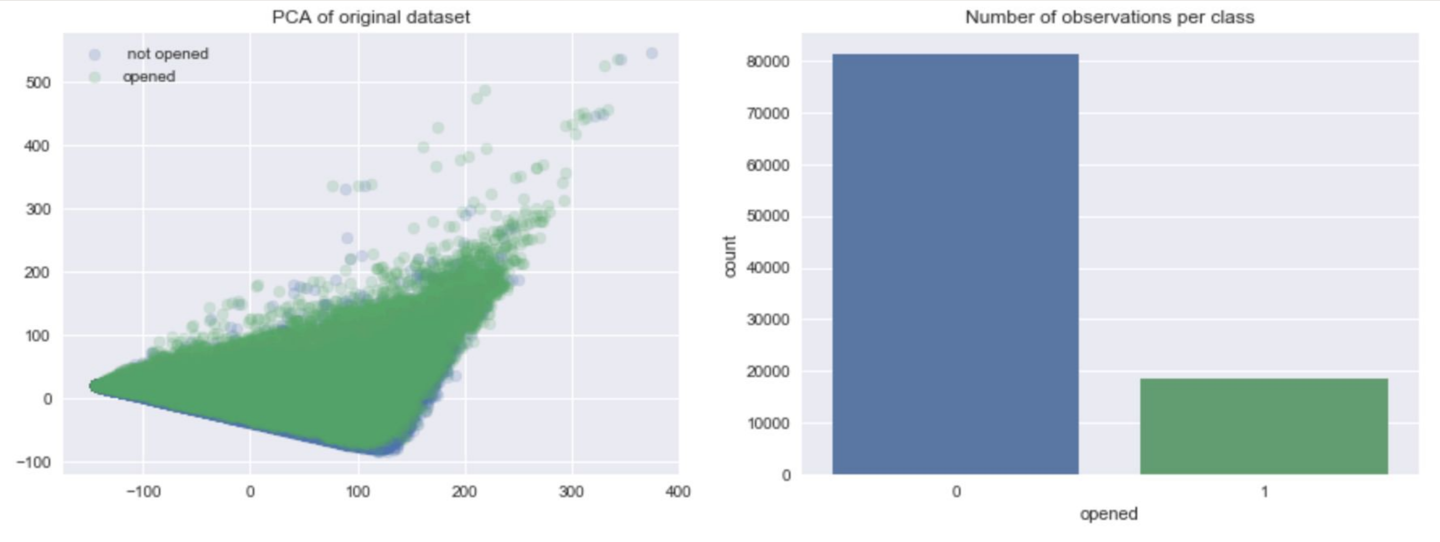
UserInfo
riid
aq_dt
Sub_source

Sends
riid
Campaign_send_dt
Opened

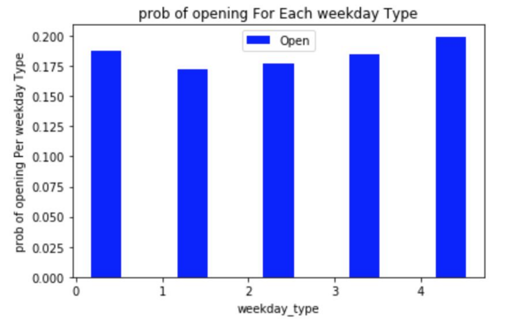
Campaign Types
Launch_id
camptype

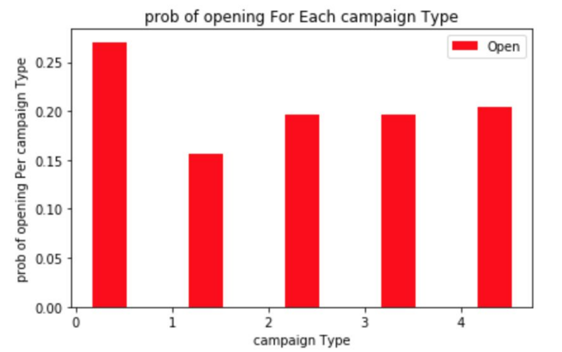
Events
riid
event_captured_dt
event
Cnt

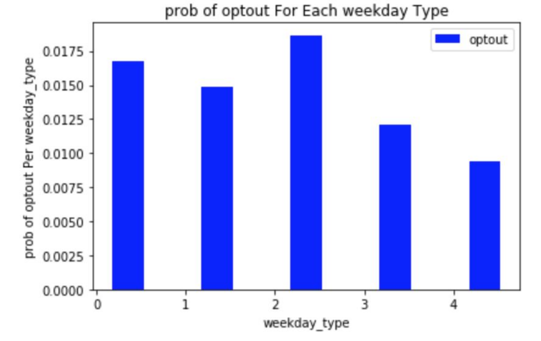
Class imbalance in dataset

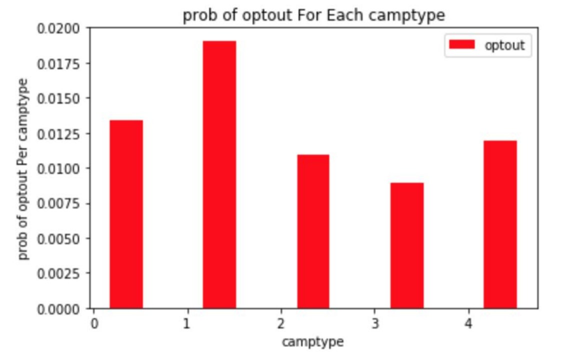


Few campaign types and weekday types have high open rate and optout rates

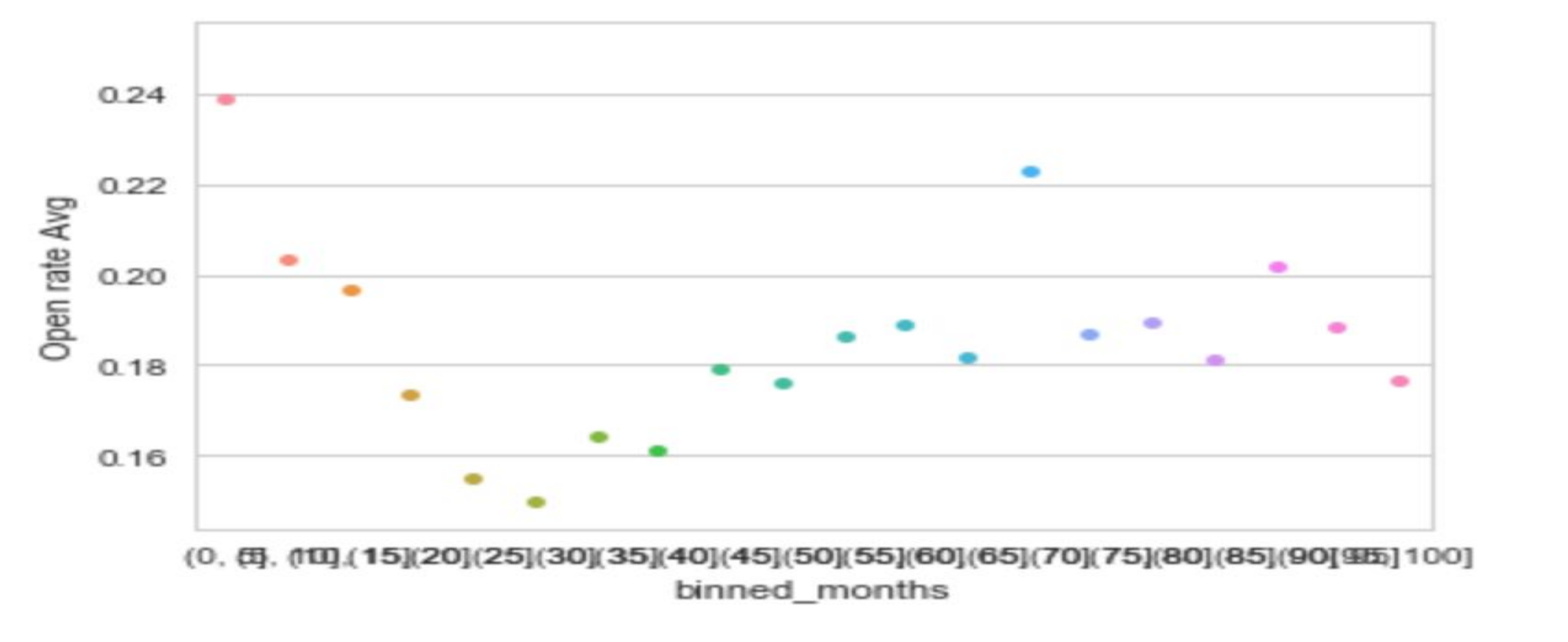




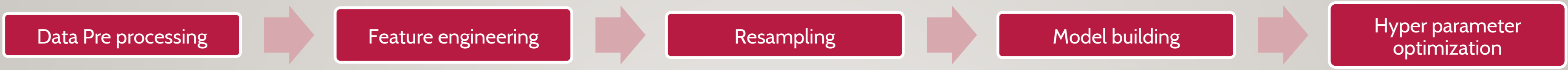




Subscription period vs Avg open rate



Methodology



Data Pre processing

Feature engineering

Resampling

Model building

Hyper parameter optimization

Class imbalance is observed in “Opened” and “Optout” variables

Objective : High precision on “Not opened” class and High recall on “Opened” class

Resampling techniques used to resolve class imbalance

Oversampling: RandomOverSampler , SMOTE , ADASYN

Under sampling: RandomUnderSampler(), NearMiss1, NearMiss 2, TomekLinks(),EditedNearestNeighbours

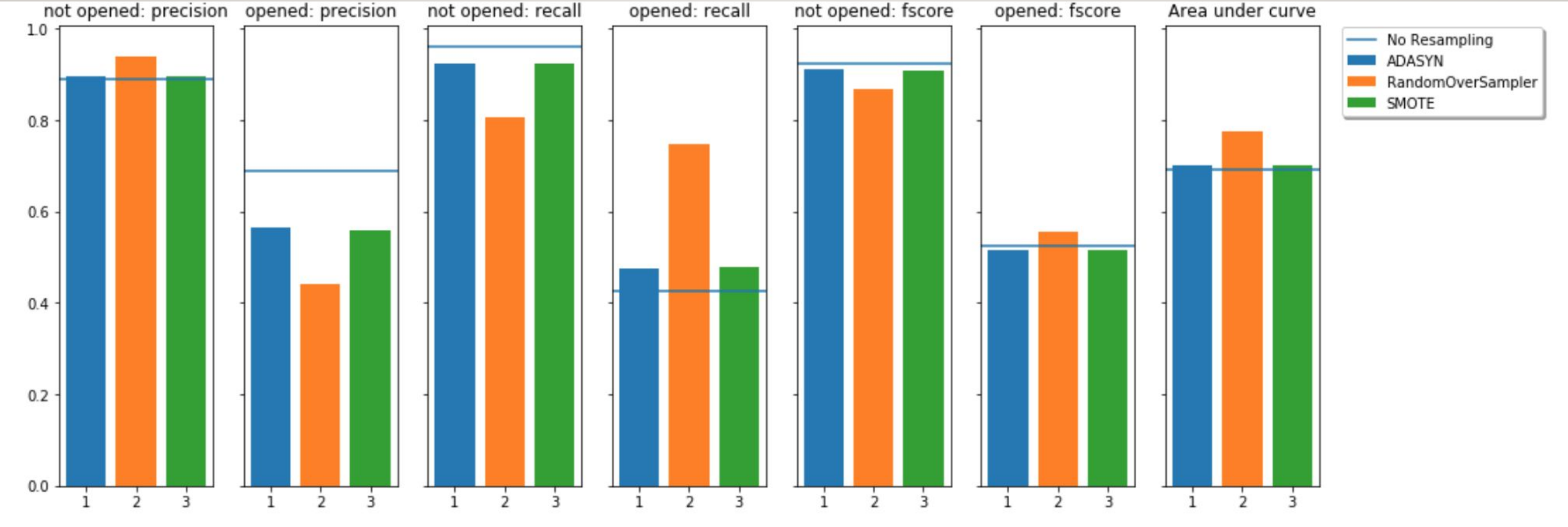
Combination: SMOTEENN , SMOTETomek

Feature engineering

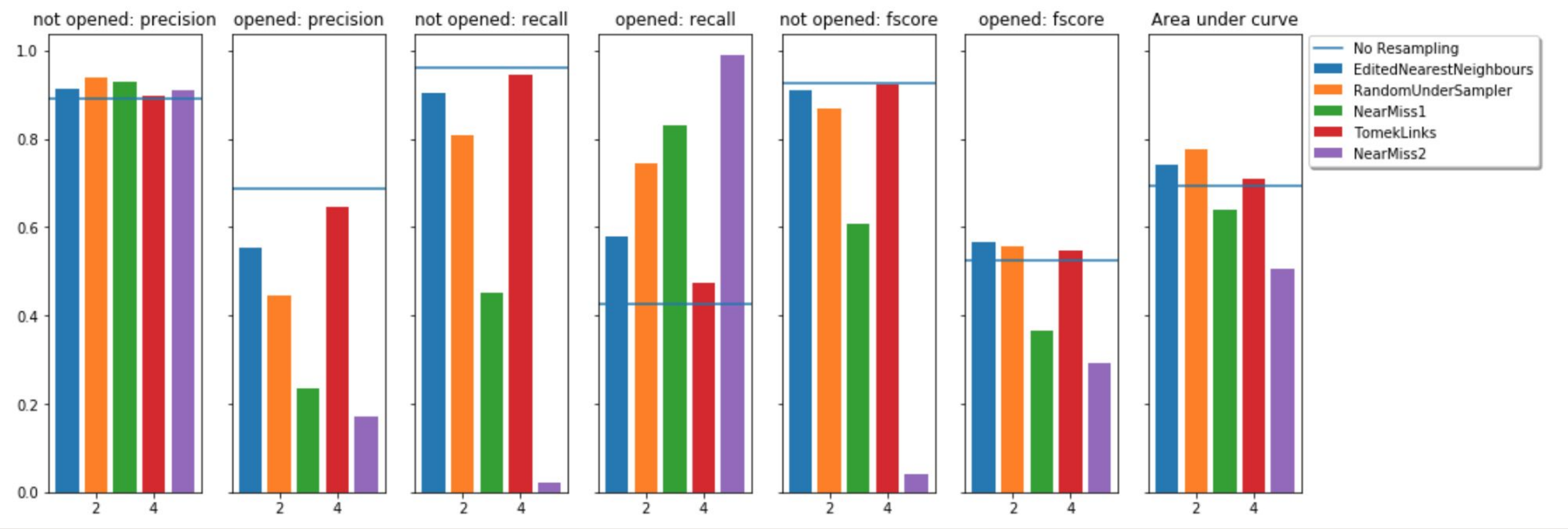
Extracted the below features

1. Tenure
2. Weekday Type
3. Cumulative Clicks, Opens, Sent, Online purchase @ user level
4. Time difference between the previous campaign and the current
5. Cumulative Opens @ user, campaign level
6. Min, max, mean and standard deviation of the mail sent time
7. Total # of mail campaigns per user ID
8. User mailing Recency, Frequency
9. Target encoding of riid with ‘opened’
10. Month of the acquired date

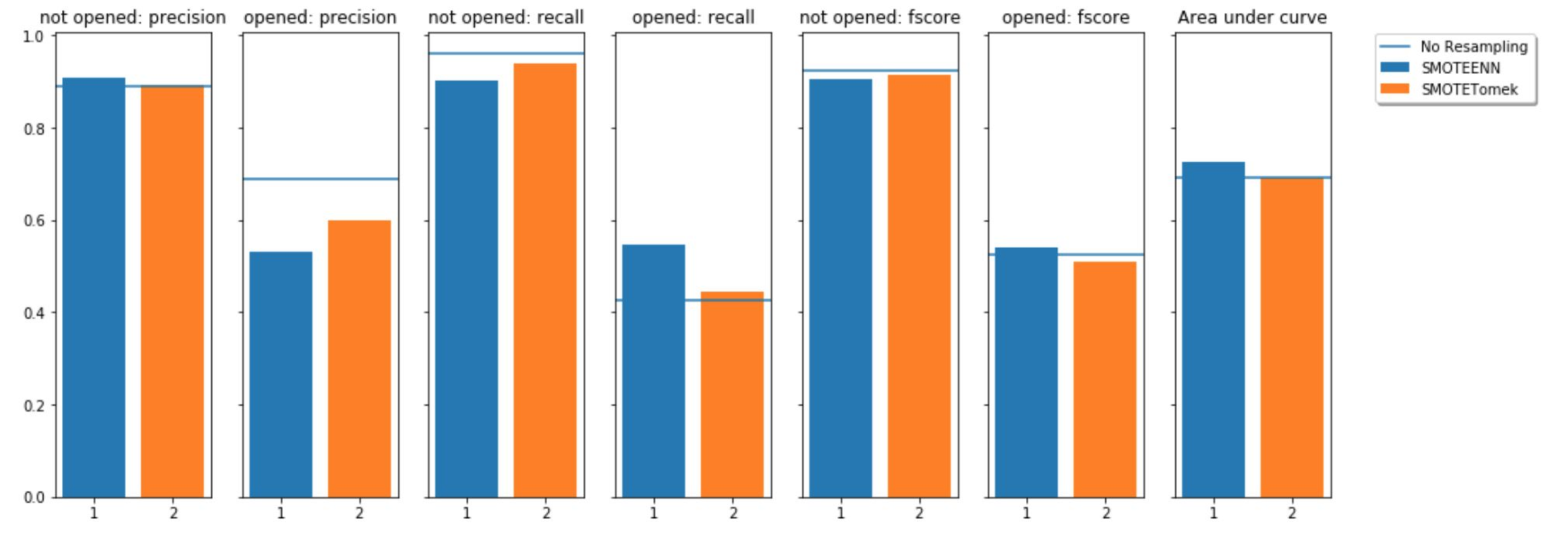
Xgboost Model Performance for predicting “Opens” with Oversampling



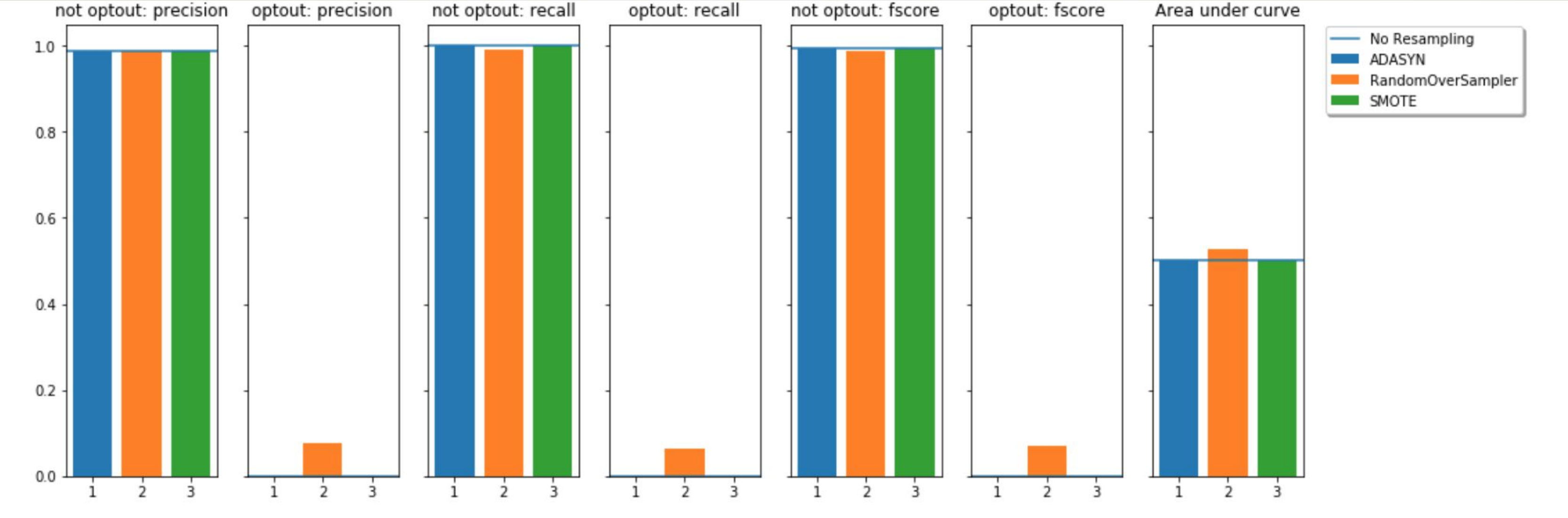
Xgboost Model Performance for predicting “Opens” with Undersampling



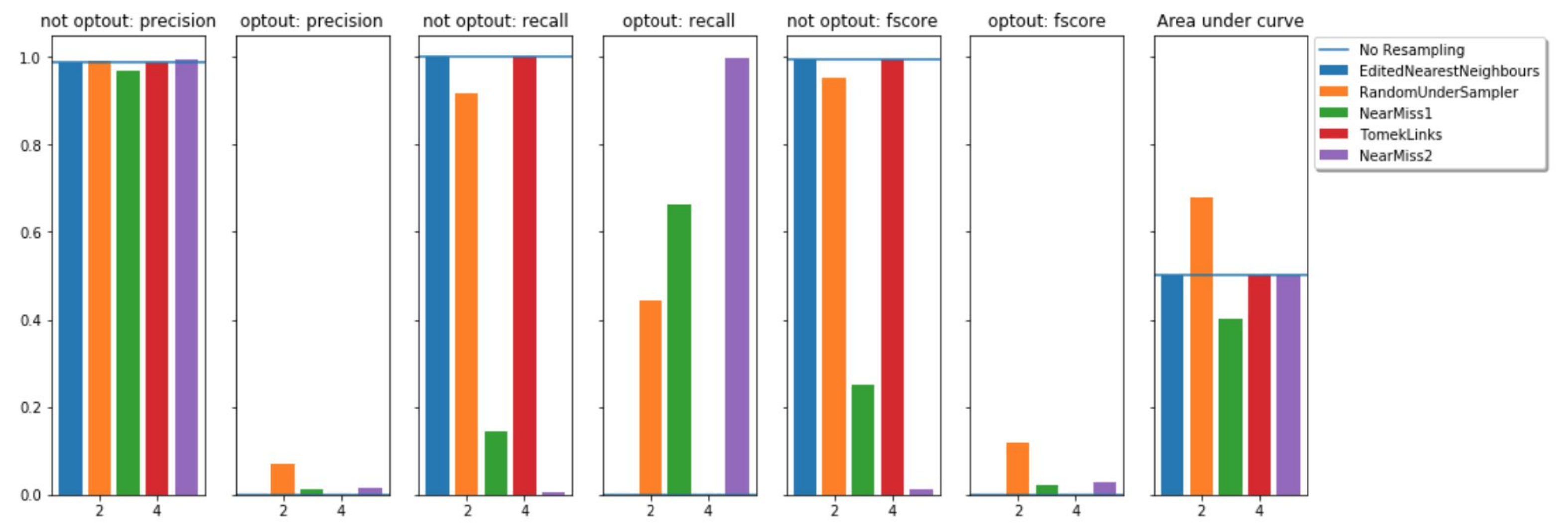
Xgboost Model Performance for predicting “Opens” with Combination



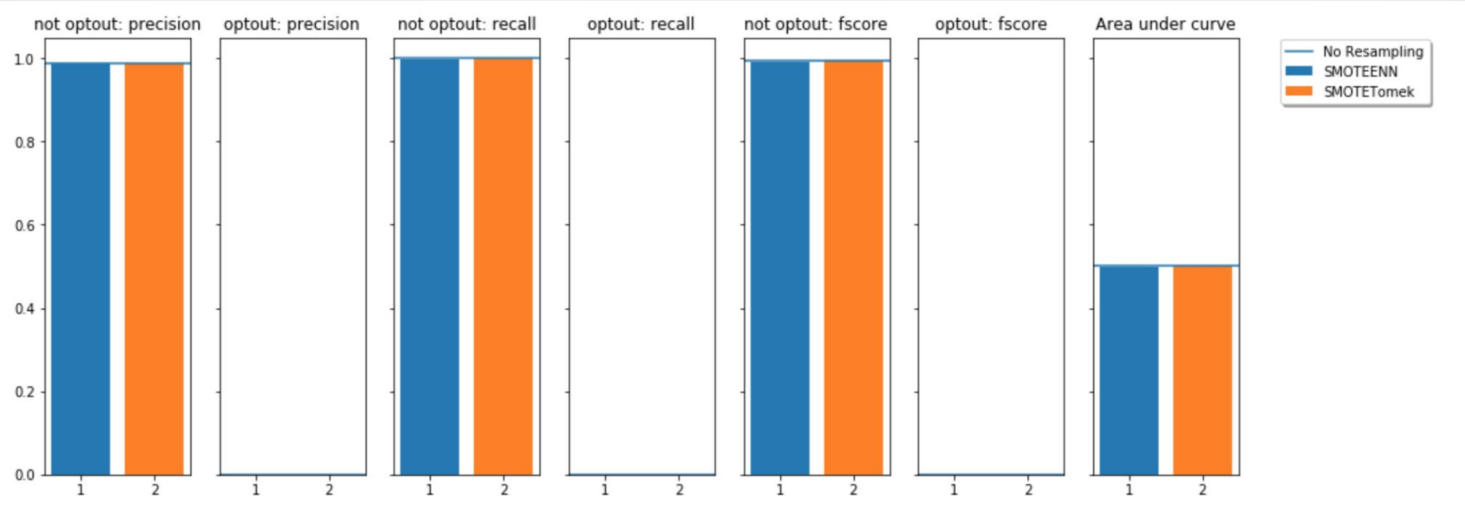
Xgboost Model Performance for predicting "Optout" with Oversampling



Xgboost Model Performance for predicting "Optout" with undersampling



Xgboost Model Performance for predicting “Optout” with combination



Insights from data exploration

The Opens table sampled users and Sends table sampled users have very few customers in common

The weekday type impacts the open and optout rates

Different campaigns have different response rates in terms of opening/clicking/optout rates

The longer the subscription period, the Avg open rates are high. However, too long tenure is opposite

The open and optout rates are varying based on Sub source type and launch dates

Some of the user information is not right

Future Improvements given more time and computational resources

Techniques

- K-fold validation in Xgboost
- XGBfir package for including 2nd order and 3rd order interactions variables
- Neural network
- LSTM,RNN for sequential predictions (Open → Unsubscribe)
- Hyper parameter tuning using hyperopt package
- Ensemble : Logistic regression , Xgboost and Extra trees models
- Using t-sne features based on unsupervised clustering of users

Data for extra feature engineering

- Campaign mail sent time (Hour, minute, seconds)
- Title of the mail
- Content of the mail : Extract word2vec features, Using tf-idf unigram,bi-gram,tri-gram
- Campaign information : what was it aimed for ? More attributes
- Cosine similarity bw subject and body
- calculate features like mean send time be two mails

A/B testing plan for the model

Given user, campaign type and date, we know the probability of opening a mail

Model assesment

A : send mails to customer recommended by the existing model
B: send mails to customer recommended by the New Model

Metric to measure: Click through rate, open rate, online purchase spend

Check statistical difference after maintaining everything else the same and see the lift in the Click through rate, open rate, online purchase spend and decide if the model is effective than the previous model

Action Items for email marketing manager

- Right content personalized to the right customer at the right time is what drives engagement rates and minimizes optout rates

- Implement the model and do A/B test before rolling it over to all customers/campaigns

- Log/Collect more data which will improve the email effectiveness

- Integrate other activity of the user also into the model. Set up the system so that
- It can be fed into the email sending pipeline

- To optimize for right content: log/collect Title of the mail ,Content of the mail , links, images in the mail

- Campaign information such as what is it about ? what is it promoting ? which item? category ?

- Customer details: Other transaction details on the website to link both of them to have a better understanding about the customer