Finding Potential Customers to advertise Carribean Holiday Tour

Methodology

Business Understanding

Caribbean Holiday Operator looking to maximize customer conversion given limited marketing budget

Data Understanding

Data is produced using an online quiz where images, corresponding to a particular tag, are shown to the users and user has to select image(s)

Data Preparation

Data cleaning and clustering to identify the characteristics of the users

Modelling

Models testing (Logistic Regression with Lasso Penalty, Random Forest, Random Forest with Hyperparameter Optimization, XGBoost)

Problem Statement

A Caribbean Holiday Operator is looking to advertise their trips to potential customers online.

Resource Constraints

Due to the marketing budget constraints of the client, we have to target the users with highest probabilities of conversion.

Actionable Insights

Provide insights about the characteristics of the users who have bought the tour in the past

Model Predictions

Select the best model and predict the conversion probabilities of the users

Data Information

Different types of tags in taxonomy - 670

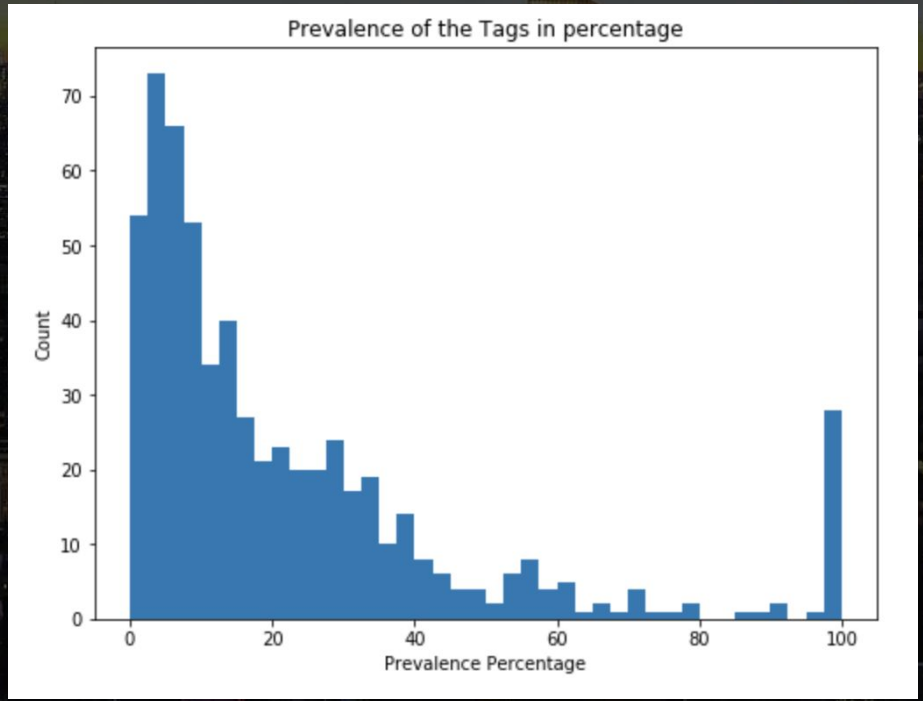
Total tags in the data - 1277 (either _A or _P)

Total attempts for the users - 7495

Users taking the quiz more than once - 104

Possibilities for a tag:

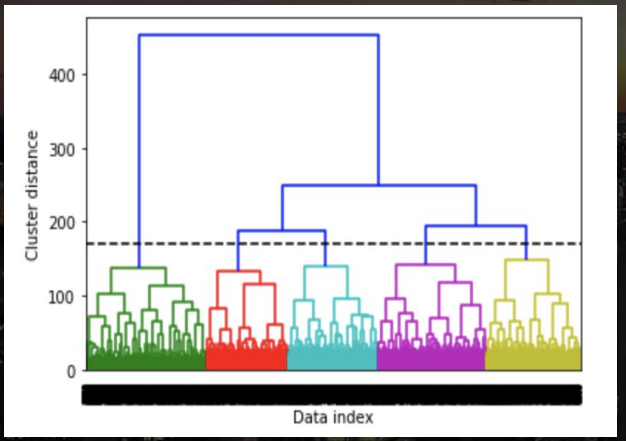
- 1 if option was shown and user chose the option
- -1 if option was shown and user didn't choose the option
- 0 if users was not shown the option
- Continuous



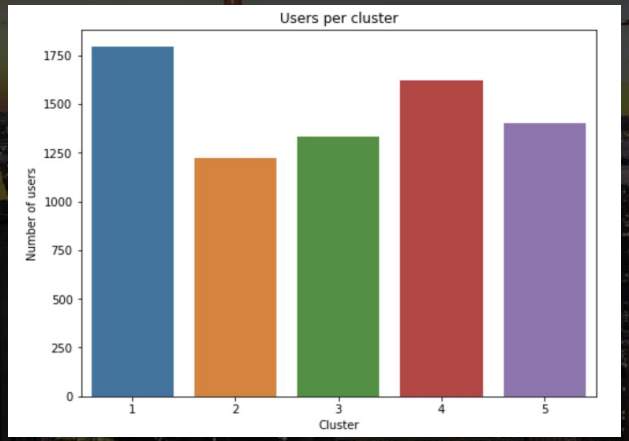
Data Preparation

01	**Tags encoding**	• 1 if user chose the option • -1 if users didn't choose the option • 0 if user was not shown the option
02	**Removing data with duplicity and errors**	• 104 users took the test more than once Considered only the last result • Removed columns marked as _A that didn't have corresponding _P
03	**Multicollinearity**	• Set the multicollinearity threshold at 0.75 • Dropped 164 columns
04	**Class Imbalance**	• Corrected using SMOTE

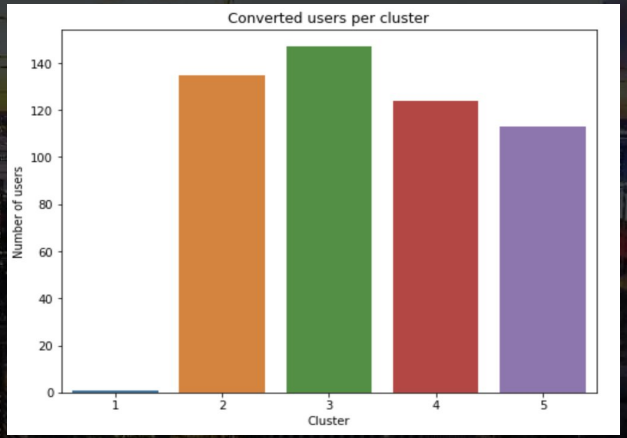
Clustering using Dendrograms



5 clusters were chosen using the dendrograms as it represents the maximum drop in cluster distances



Distribution of all the users in different clusters.

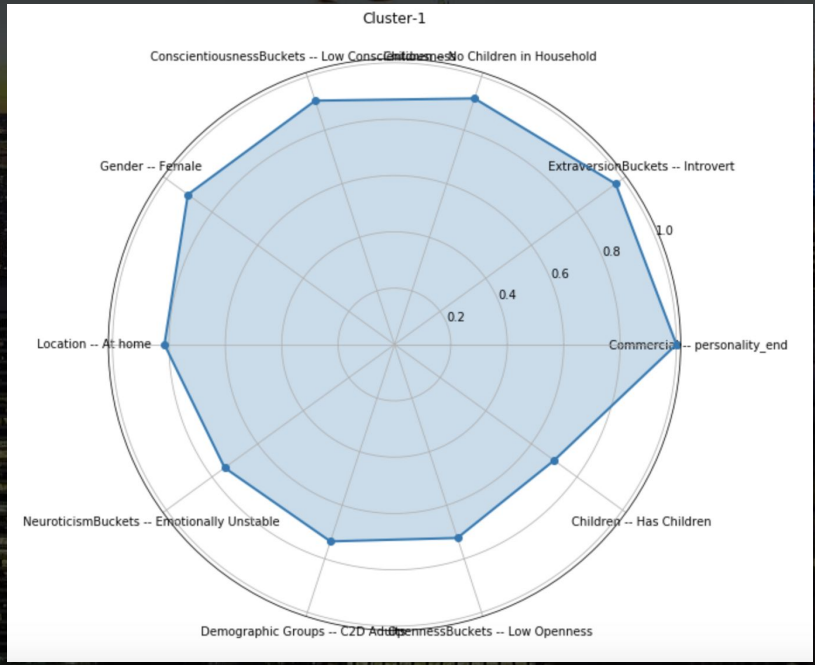


Cluster distribution of the users who purchased the tour in the past.

Cluster-1 (User Characteristics)

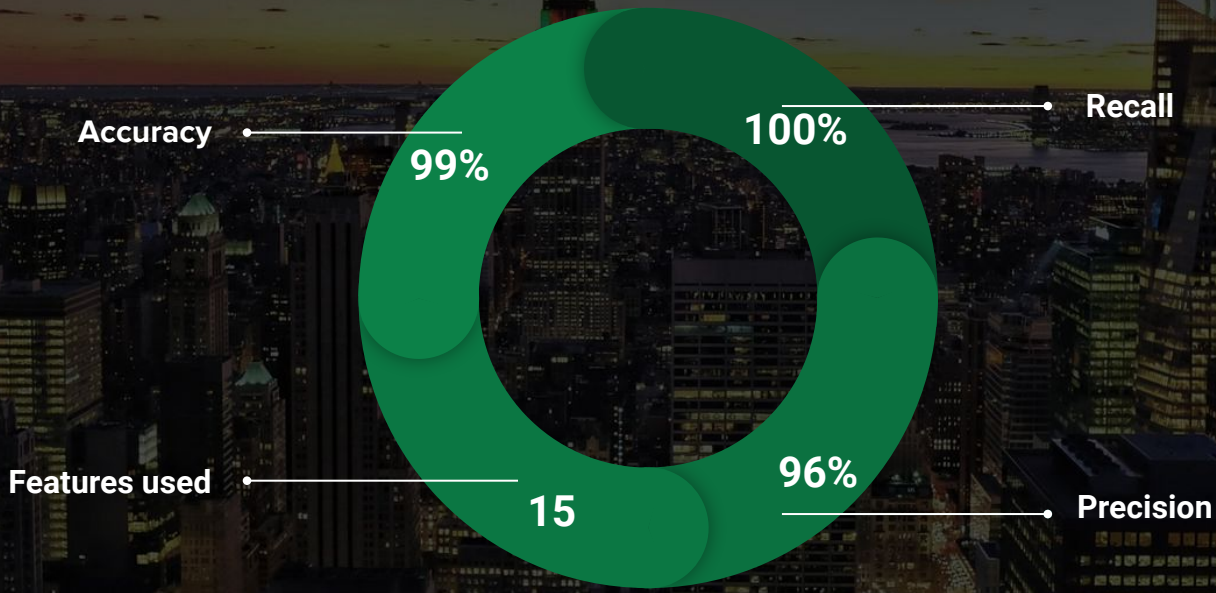
	Label	Cluster-1	Label-Last Two
33712	Taxonomy5 § Editorial Segments § Commercial § ...	1	Commercial -- personality_end
11645	Taxonomy6 § MediaBig5 § MediaBig5Buckets § Ext...	0.969883	ExtraversionBuckets -- Introvert
30399	Taxonomy4 § Demographics § Children § No Child...	0.918572	Children -- No Children in Household
11642	Taxonomy6 § MediaBig5 § MediaBig5Buckets § Con...	0.910206	ConscientiousnessBuckets -- Low Conscientious...
9411	Taxonomy4 § Demographics § Gender § Female	0.905745	Gender -- Female
35959	Taxonomy4 § Demographics § Location § At home	0.815951	Location -- At home
11649	Taxonomy6 § MediaBig5 § MediaBig5Buckets § Neu...	0.744562	NeuroticismBuckets -- Emotionally Unstable
30686	Taxonomy4 § TV Audiences § Demographic Groups ...	0.735081	Demographic Groups -- C2D Adults
11639	Taxonomy6 § MediaBig5 § MediaBig5Buckets § Ope...	0.722253	OpennessBuckets -- Low Openness
9418	Taxonomy4 § Demographics § Children § Has Chil...	0.699944	Children -- Has Children

As we see from the user characteristics, users who bought the tour in the past don't possess combination of these characteristics.



Base Model & Feature Selection

Logistic Regression with Lasso Regularization (0.03)



- Results seem too good to be true.
- Problem: Model was trying to predict using the columns where option was not shown in most of the cases.
- Removed the columns where options was not shown in more than 50 percent of the cases

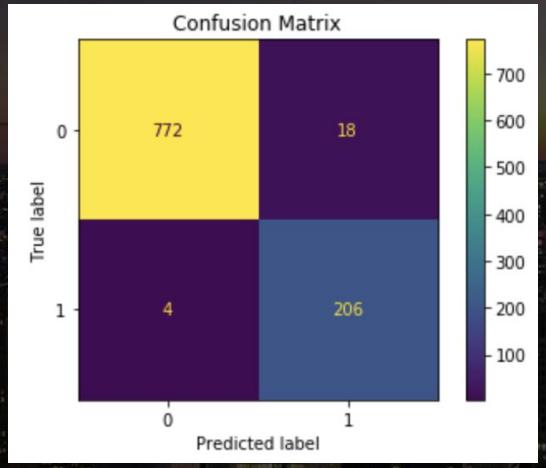
Models Tried

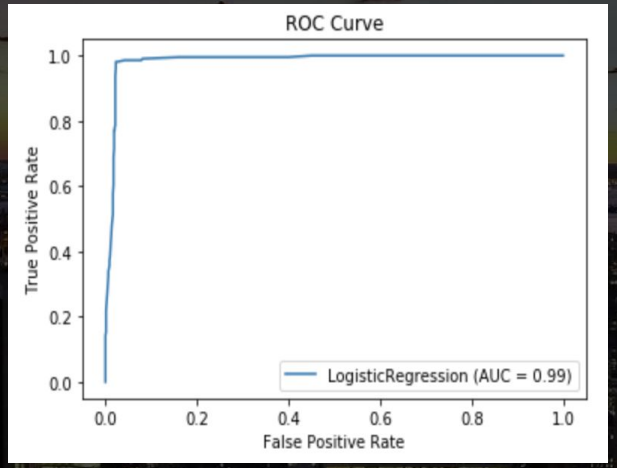
 98%	• Logistic Regression with GridsearchCV (Lasso Penalty)	• 98% Accuracy • 92% precision • 98% recall
 95%	• Ensemble - Random Forest	• 95% Accuracy • 92% precision • 81% recall
 98%	• Ensemble - Random Forest with hyperparameter optimization	• 98% Accuracy • 92% precision • 98% recall
 98%	• XGBoost	• 98% Accuracy • 92% precision • 97% recall

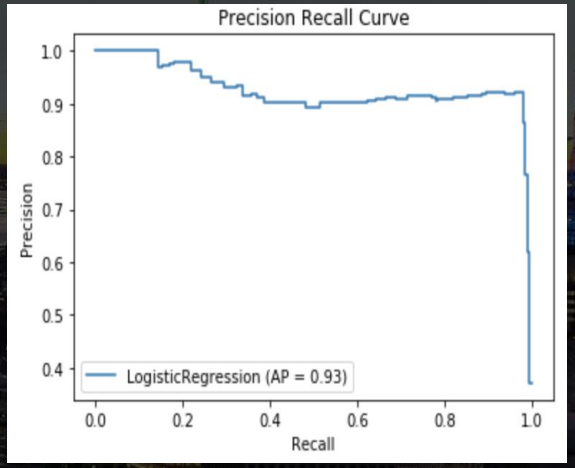
Choosing the best model:

- Computationally cheap: Logistic Regression
- Number of features used: 9 in Logistic Regression
- Priority to recall because the cost of loss of potential sale outweighs the cost of marketing: Logistic Regression

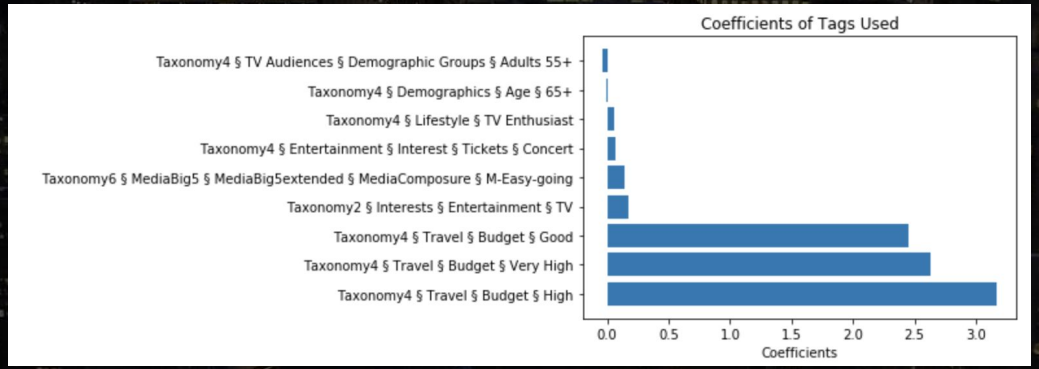
Results







Features importance tend to be coherent with primarily having a high budget, being in an older demographic group coupled with high media consumption (ie. tv, concerts etc). That said, given the limited marketing budget, it is suggested to target the users for which the predicted probability is very high.



Q&A