Predict the default behavior of small personal loan applicant

Business Content Overview

- Rising market of small personal loan

- Leveraging data science and machine learning technology to capture the rising market

Data Acquisition and Description

- From the Telecom company's internal data database; retrieve by SQL

- Three tables in total:

  Phone Call records

  Application downloads

  Label of Defaults

Data Wrangling for Phone Call Records

- Missing value(NA) : drop two columns which contain the majority of missing values but similar information is contained in other columns; then remove all of rows that contain missing values

- Aggregate data of same user (same phone number) :

  Sum the call durations for each user according to the call types (calling/called/call transfer)

  Sum the call duration for each user according to the business zone types (domestic/international).

Data Wrangling for Phone Call Records

Create a series of new features based on the data of ‘ROAM_TYPE’, 'OTHER_PARTY', 'VISIT_AREA_CODE' and 'CALLED_HOME_CODE':

1. roam ratio: use the ratio of number of roam calls/number of all (roam + non-roam) calls to represent the roam levels of each user.
2. diversity of the other party: count the distinct calling numbers and called numbers of each user to represent the diversity of user’s contacts.
3. diversity of visiting area: count the distinct visiting places of each user to represent the diversity of user’s visiting area/travel footprint
4. diversity of geography locations of the other party: count the distinct numbers of geography location for each user’s contacts to represent the geo-diversity of user’s contacts.

Data Wrangling for Application downloads

- Missing value(NA) : Drop all of rows that contain the missing values.

- Aggregate the download flux information based on application types:

  The unit time flux information is represented by download flux divided by time, that is, FLUX/DUR

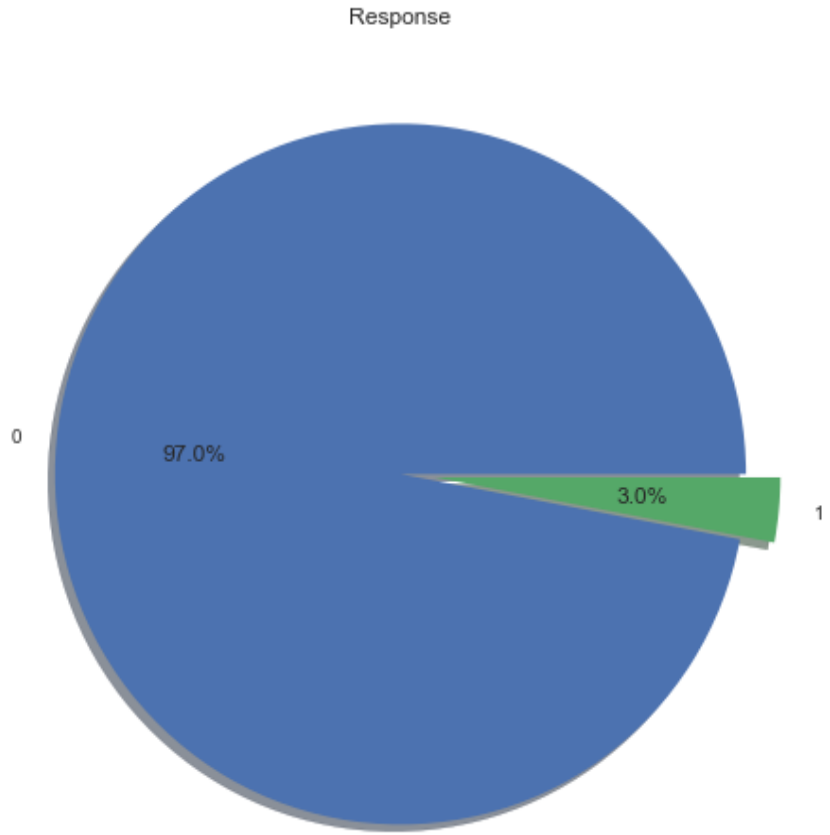
  Sum the unit time download flux based on different downloaded application types(TYPE_CODE1)

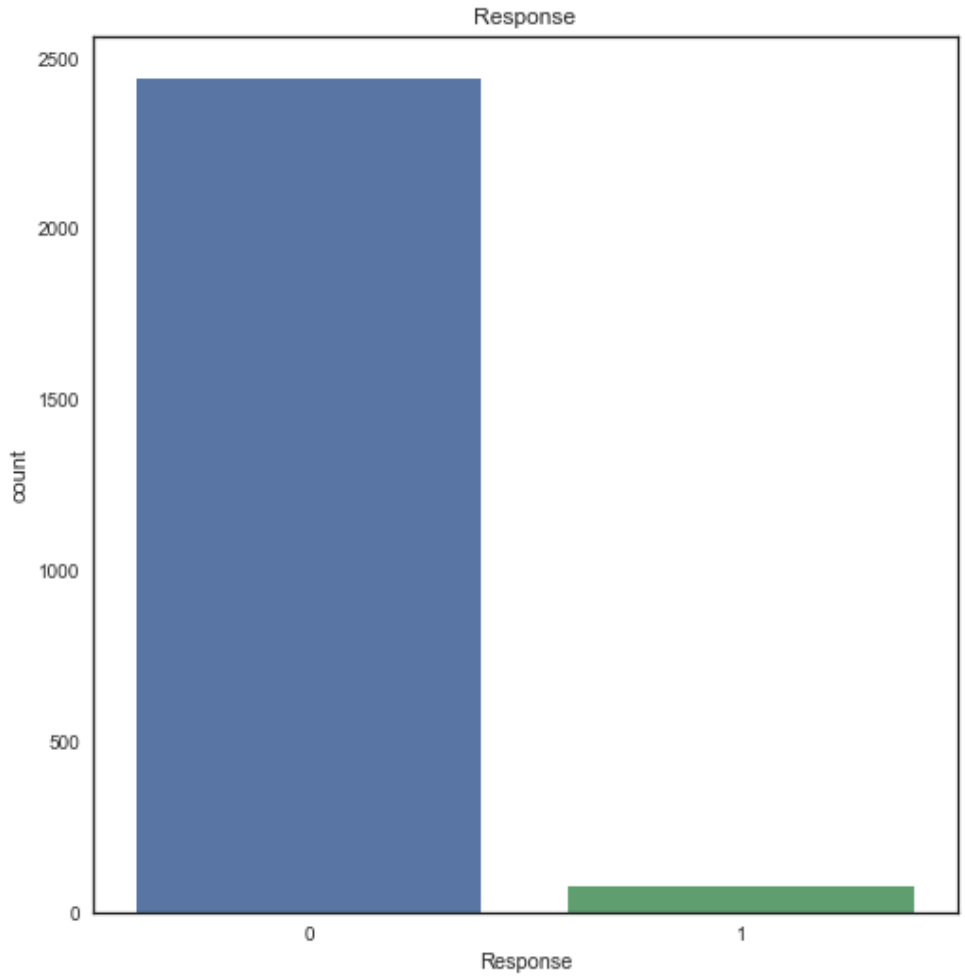
  Square root of unit time flux

EDA

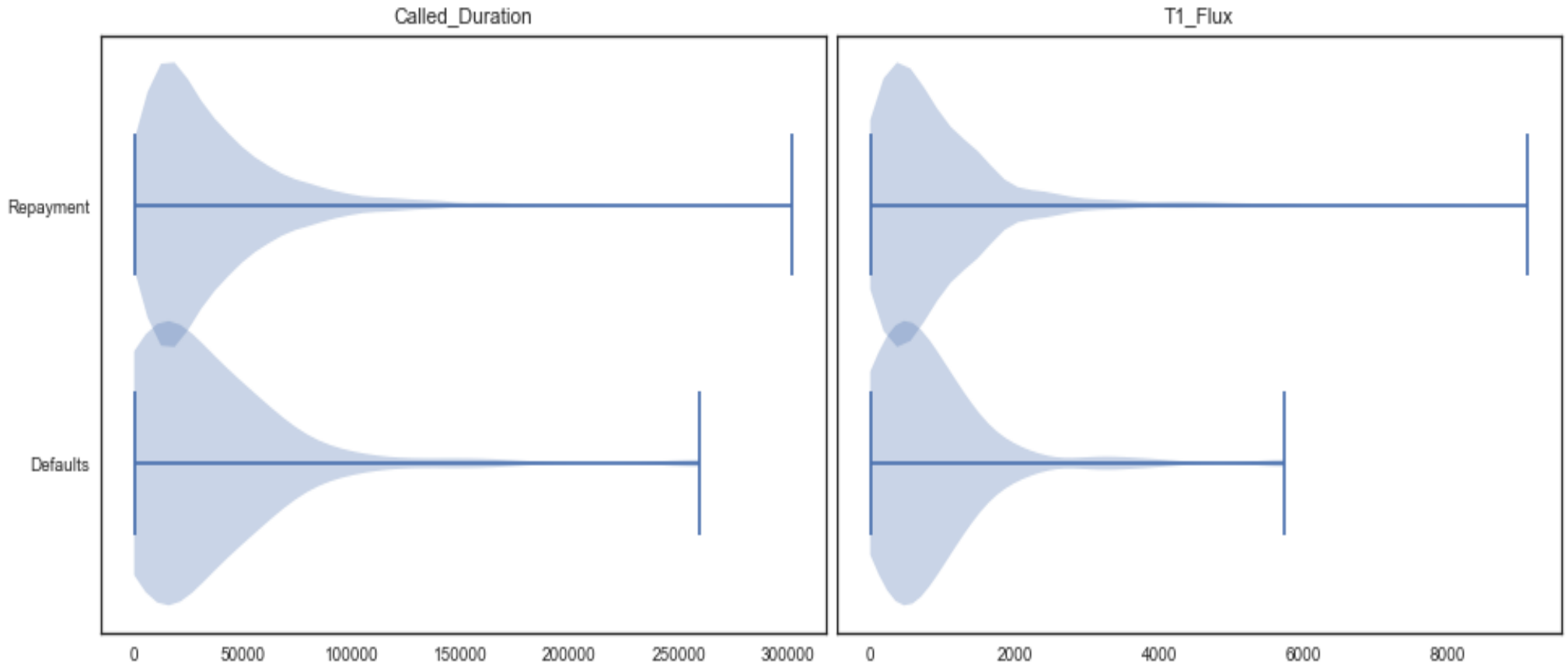
1. Look at label distribution via pie plot and histogram
2. Look at feature distributions via box plot and violin plot
3. Look at feature correlations via correlation map
4. Comparison of median number and max between two classes by bar plot
5. Test statistics for median number

Majority of applicants (97%) are not default.

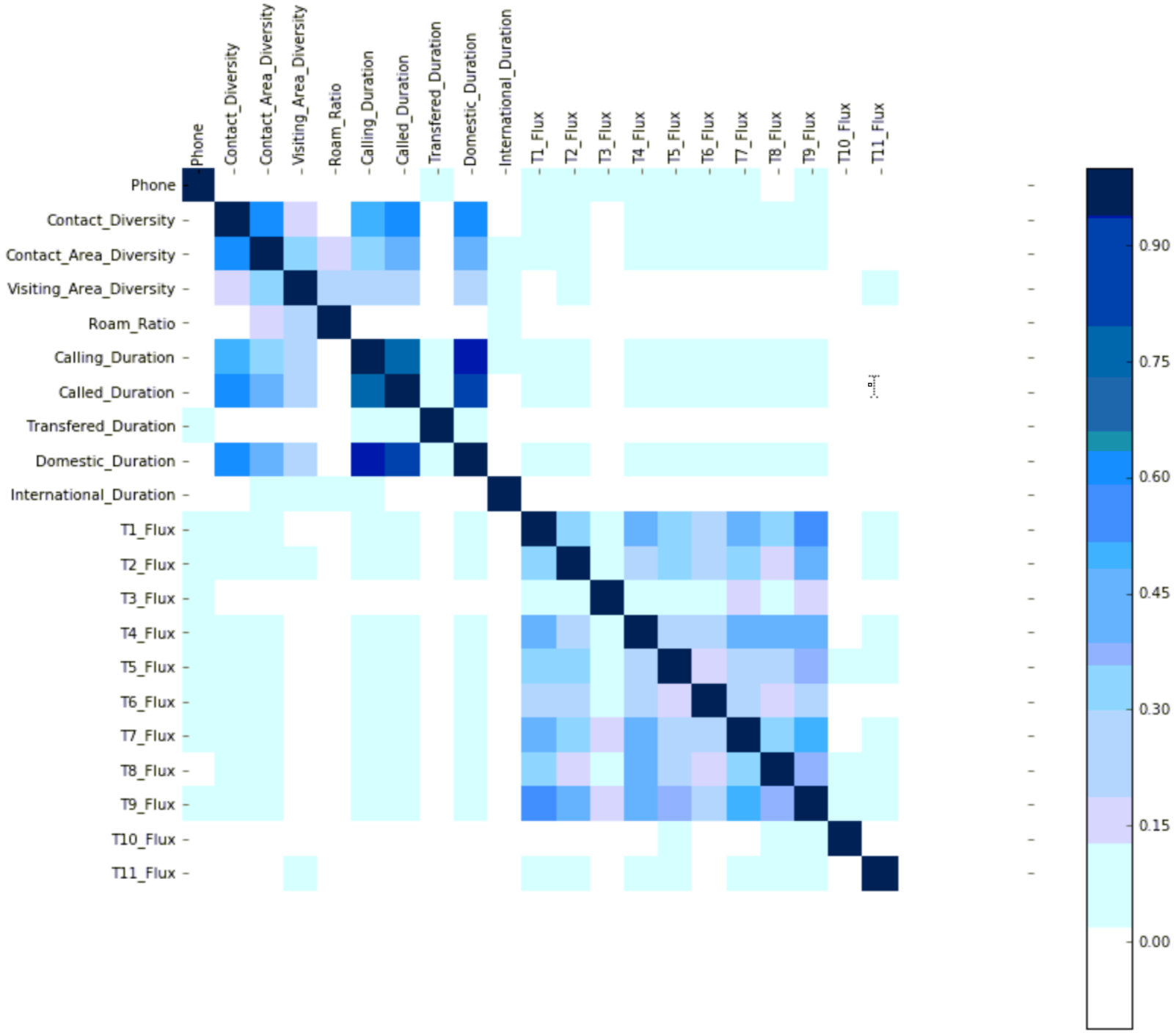


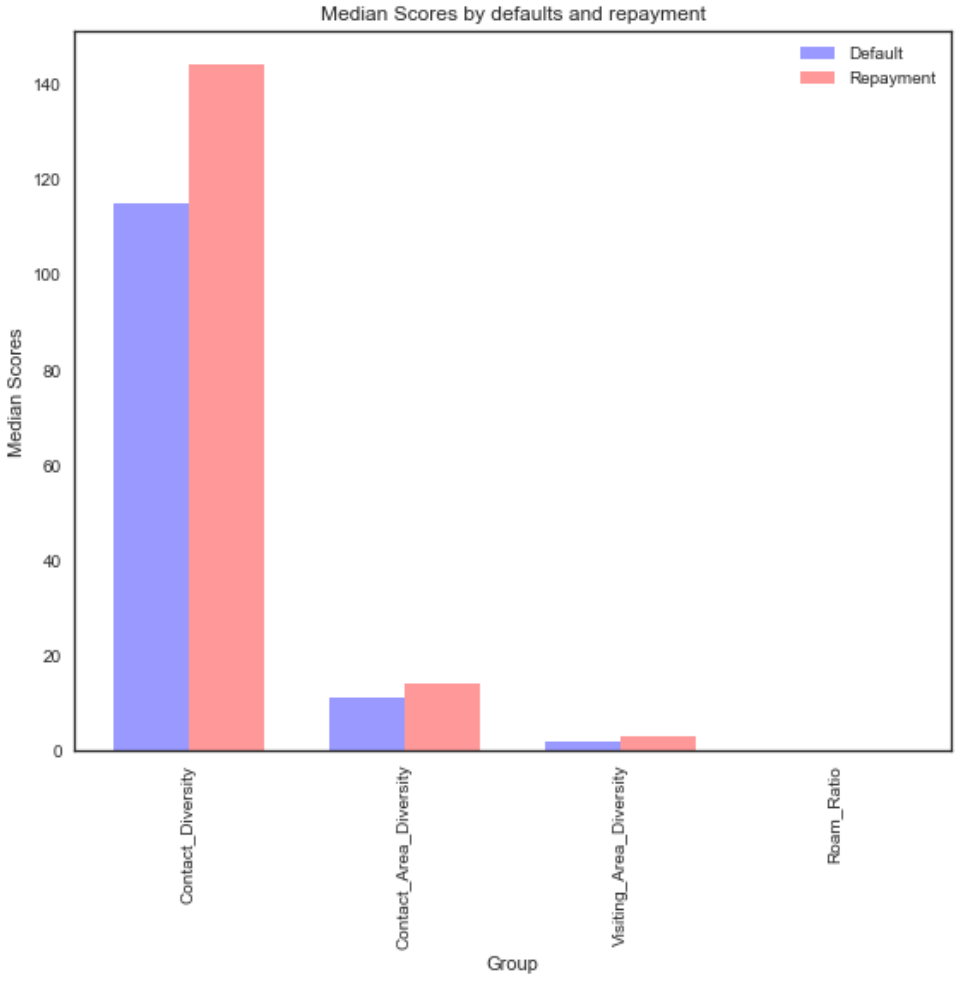


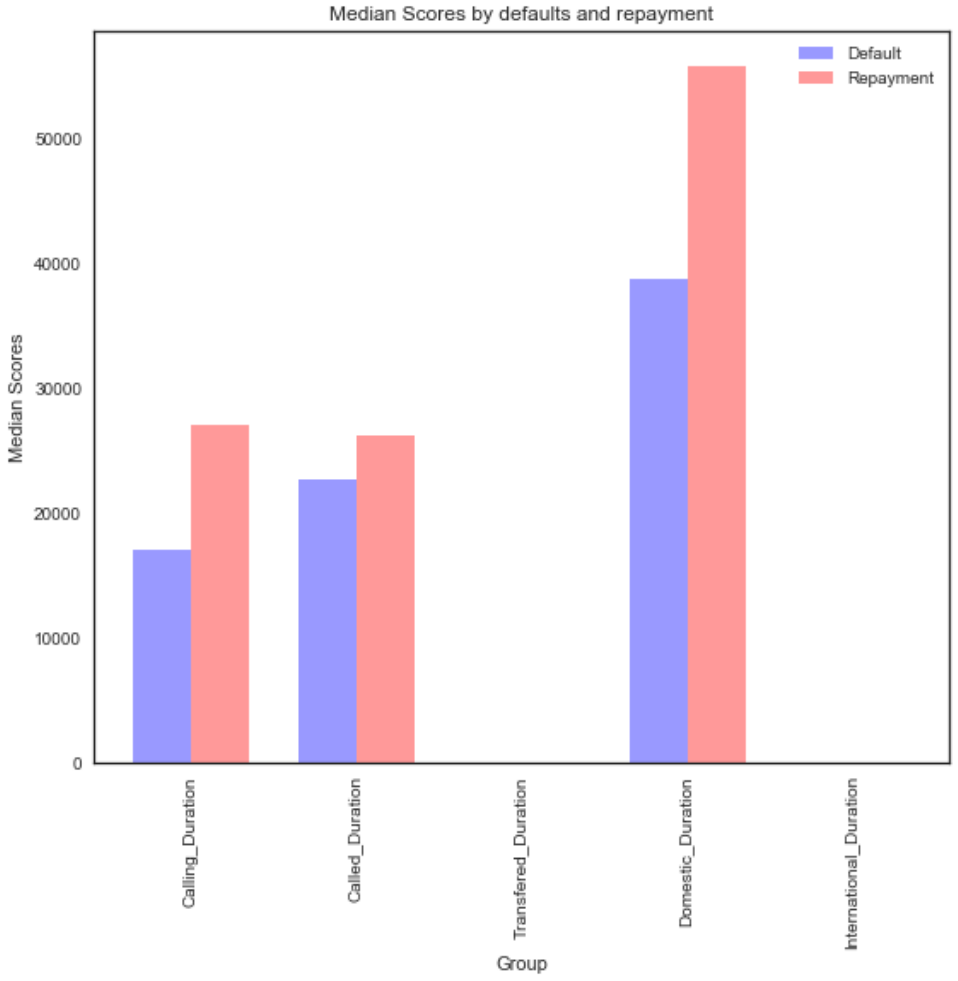
Skewed distributions for features



- Correlation between the features in Call records and those in Downloaded application is weak
- Correlation between the features among Downloaded application is weak
- Correlation between the features among Call records is strong







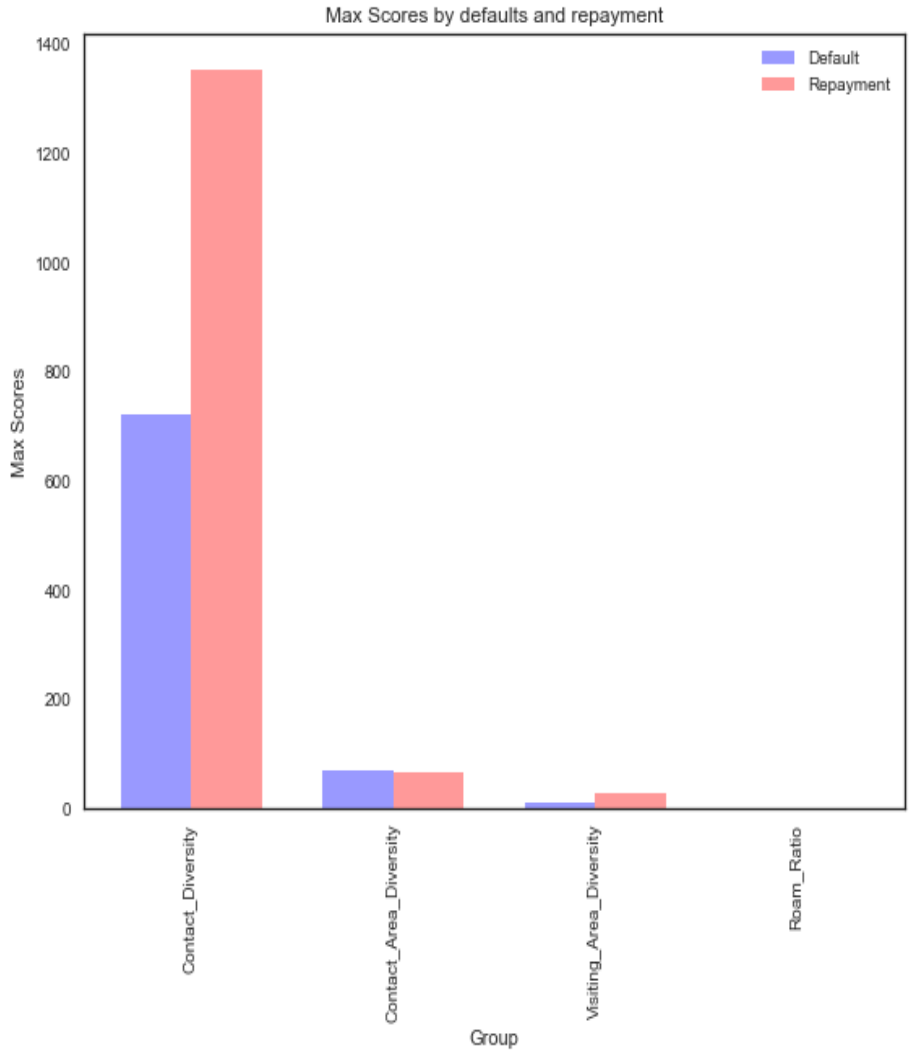
Repayment applicants

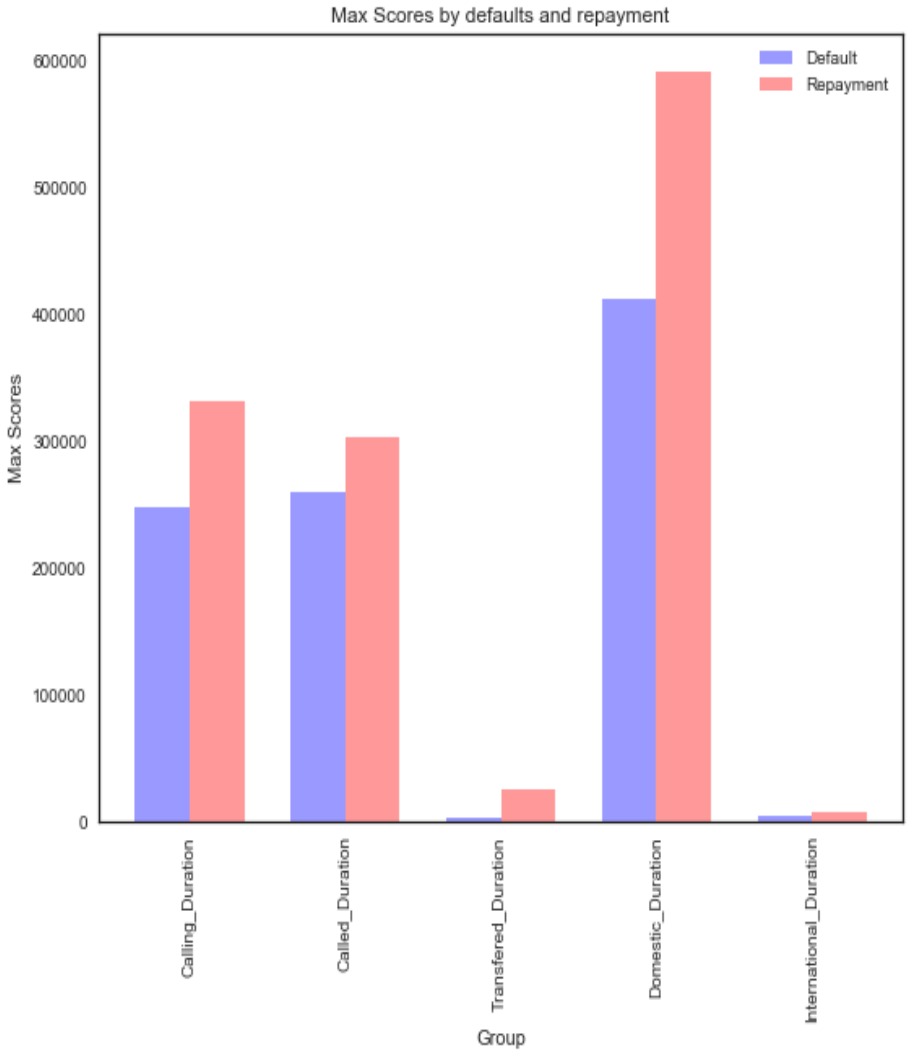
larger average number of contacts and areas diversity of their contacts.

more diverse visiting area

higher roam ratio on average

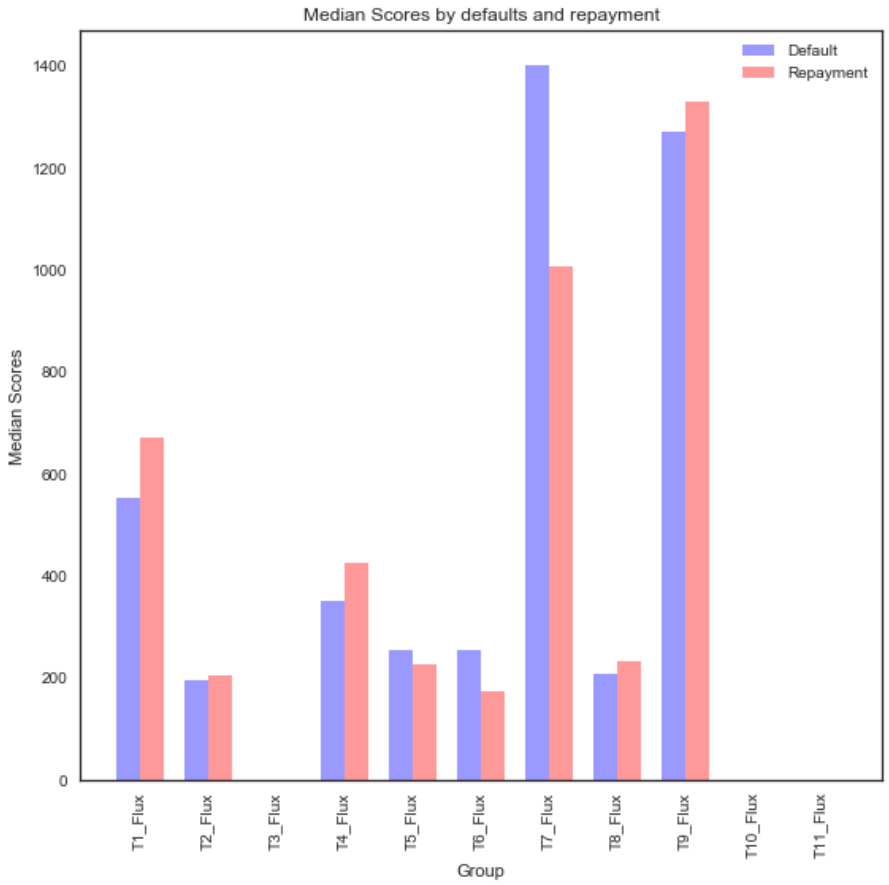
longer average calling, called and domestic call duration.

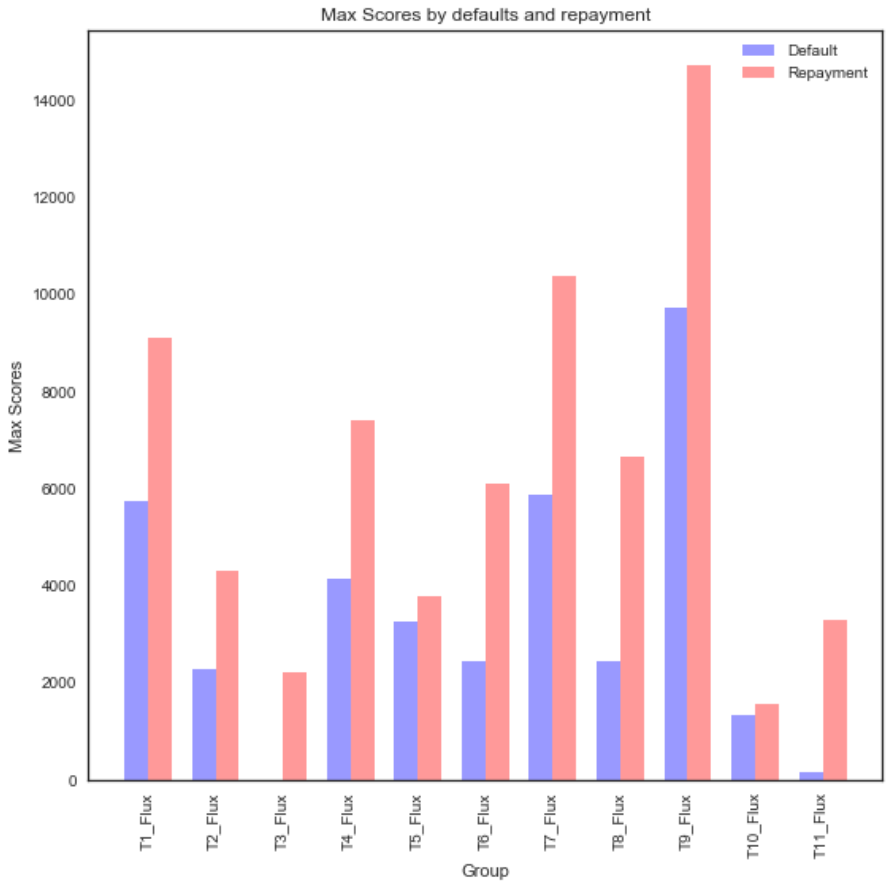




Almost the same characteristics with the median number

Except for the feature of the areas diversity of applicant contacts, showing a reverse that the repayment applicants have smaller maximum areas diversity of their contacts





- Repayment applicants have higher average downloaded unit time flux in most of categories, lower average in the categories of life service, investment management and news
- Default applicants, the investment management applications cost the highest average download unit time flux; repayment applicants, the entertainment applications cost the highest.
- In all maximum score of downloaded unit time flux, it shows that repayment applicants have higher score in all of categories.

Test statistics for the difference of median number between two classes

- For most of features, the difference of median number is not statistically significant and we cannot reject the null hypothesis

Machine Learning

1. Deploy machine learning models
2. Handle imbalanced data
3. Change the Cutoff/Threshold values in evaluation metrics
4. Future work

Deploy machine learning models
Train results after hyper parameters

Algorithms	Accuracy	AUC score
Random Forest	0.973880	0.642456
Logistic Regression	0.972748	0.603004
Gradient boosting	0.973880	0.661304

Gradient boosting gives best performance score

Handle imbalanced data
Train results on different data samplings

Algorithm	Without sampling		Oversampling		Undersampling	
	Accuracy	AUC score	Accuracy	AUC score	Accuracy	AUC score
Random Forest	0.972181	0.567206	0.972150	0.995681	0.607143	0.624362
Logistic Regression	0.972678	0.608185	0.701841	0.768425	0.598214	0.571429
Gradient boosting	0.972181	0.635041	0.948644	0.990406	0.562500	0.650510
K nearest neighbour	0.972181	0.486633	0.828310	0.917277	0.571429	0.587372

Balanced data set gives better training results.

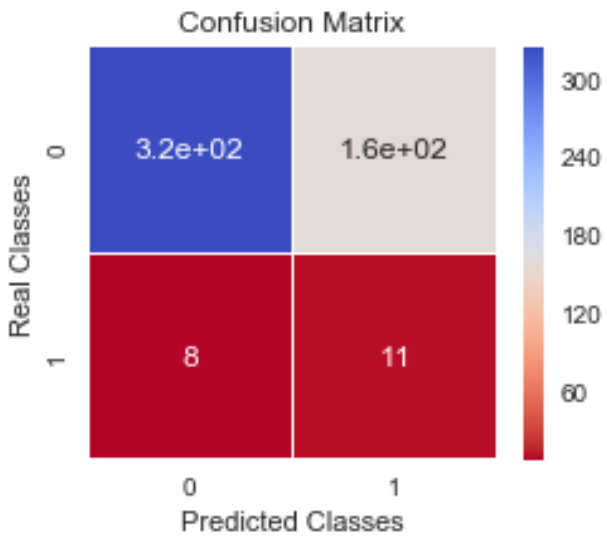
Oversampling is much better than undersampling

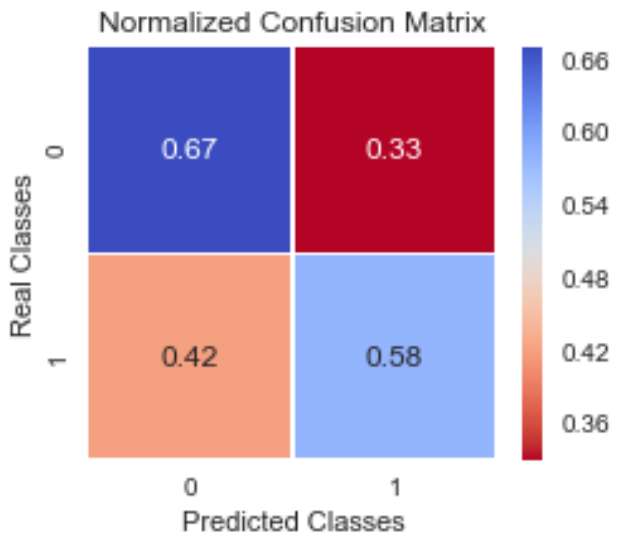
Test results on different data samplings

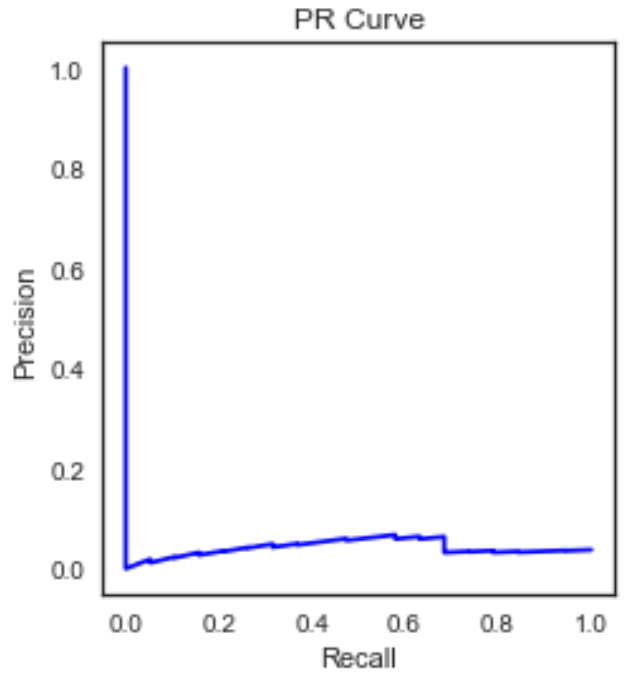
	Without sampling	Oversampling	Undersampling
Algorithm	AUC score	AUC score	AUC score
Random Forest	0.52300596853	0.51465002713	0.582528486164
Logistic Regression	0.566142159523	0.592837764514	0.53130765057
Gradient boosting	0.477590884428	0.516115029843	0.525881714596
K nearest neighbour	0.491264243082	0.507650569723	0.447693977211

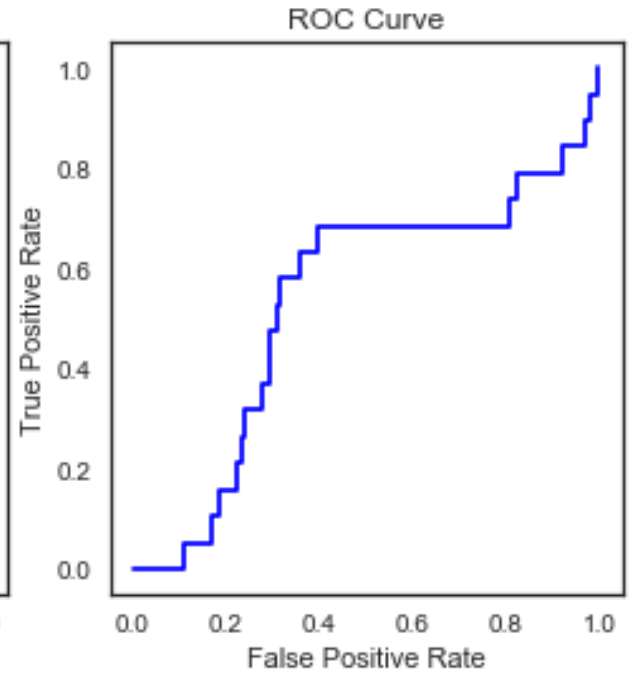
In general, oversampling performs better.
Without hyperparameter tuning, the logistic regression gives best prediction performance while tree methods suffer from serve overfitting.

Confusion matrix and Evaluation metric plot



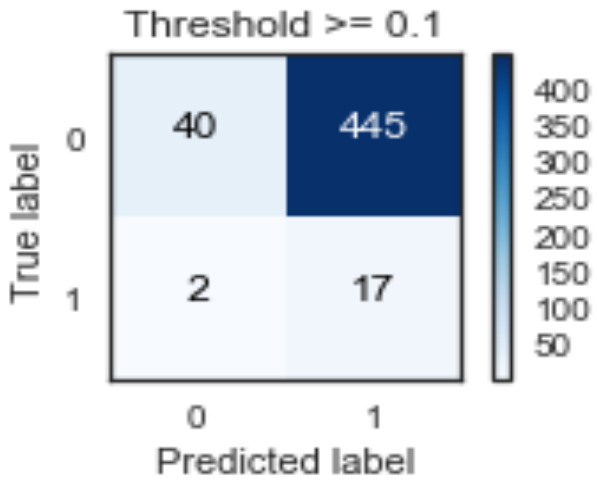




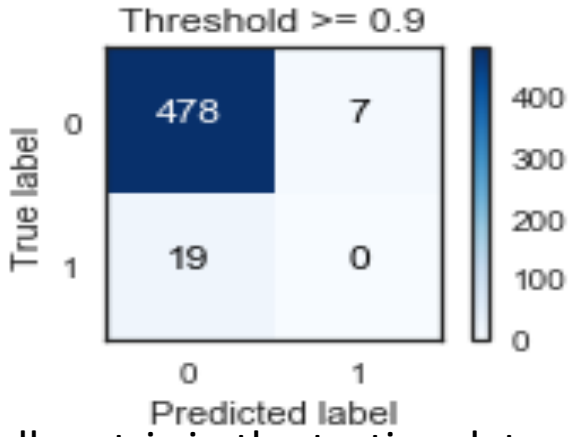


Logistic regression gives a balanced recall score on both of classes and high precision score on class 0 (repay) but low precision score on class 1 (default).

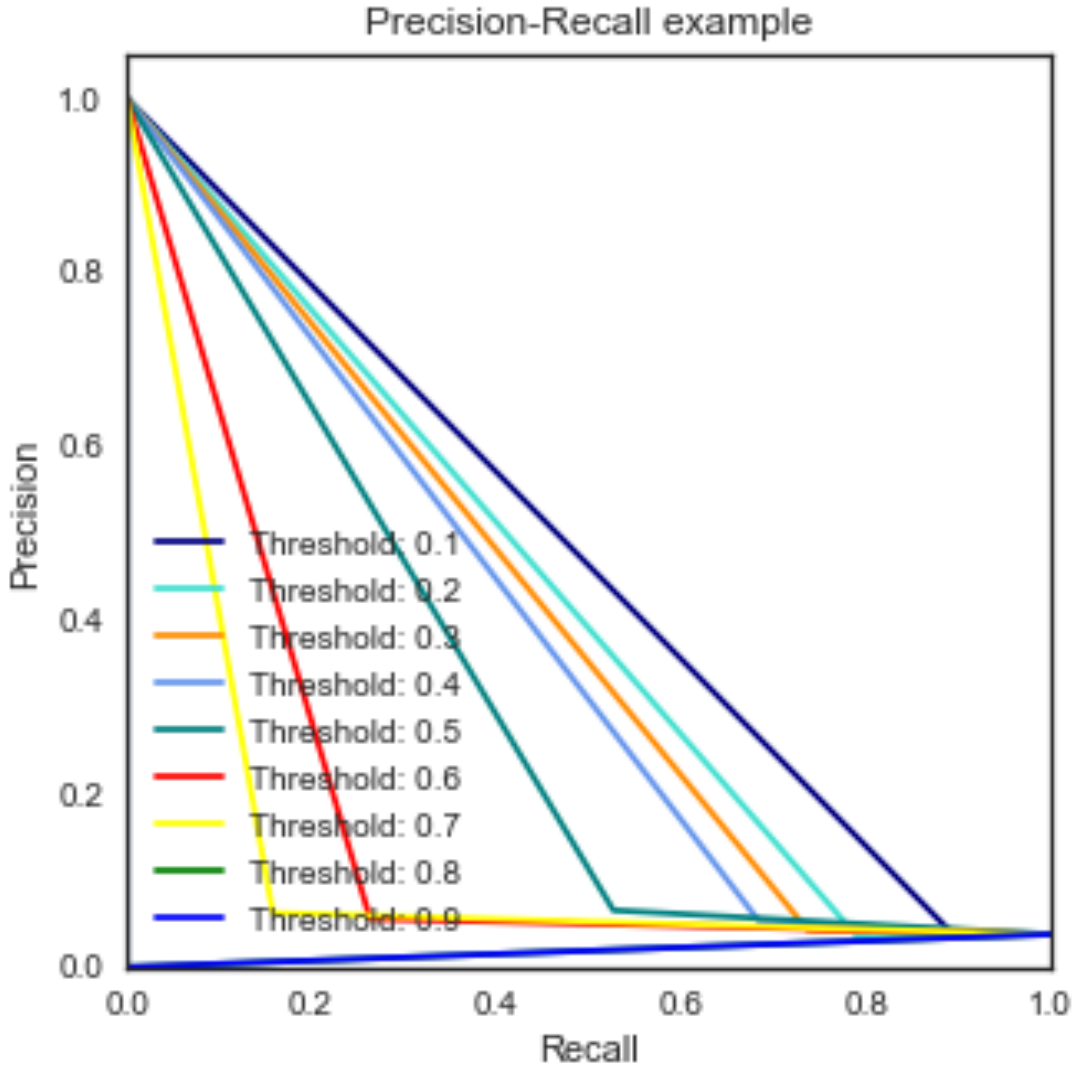
Change the Cutoff/Threshold values in evaluation metrics



Recall metric in the testing dataset is 0.89 (threshold is >=0.1)



Recall metric in the testing dataset is 0 (threshold is >=0.9)



Improve the Performance by Feature Processing

Algorithms	Feature Processing & Sampling	Test AUC score
K Nearest Neighbours	Transformed Log Transformed Feature Over Sampling	**0.6875**
Logistic Regression	Dropped Log Transformed Features	0.6840
Random Forest	Standardized Features Over Sampling	0.6415