Contents

[1. General Idea 1](#_Toc491375206)

[2. Data Exploration 2](#_Toc491375207)

[1) Initial Exploration 2](#_Toc491375208)

[2) Missing Data 2](#_Toc491375209)

[3) Correlation Matrix 2](#_Toc491375210)

[3. Linear Regression Modeling 4](#_Toc491375211)

[4. Logistic Regression Modeling 5](#_Toc491375212)

[5. Further Investigation (answer to Question 3A) 6](#_Toc491375213)

# 1. General Idea

To predict total revenue and probability that total revenue was beyond 250, we adopted three steps to complete the project:

1) Data Exploration; 2) Regression Modeling; 3) Logistic Regression Modeling.

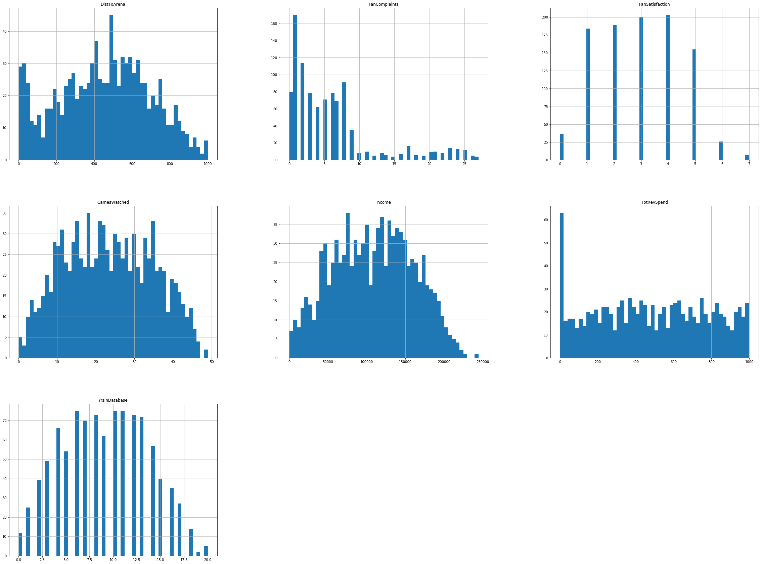
In Data Exploration, we investigated the type of variable, correlation, and other data property to gain insights about possible features.

Based on insights from Data Exploration, proper models were used to predict the total revenue.

To gain the probability that total revenue was beyond 250, the data of total revenue was transformed to 1-0 form in which 1 presented ‘>=250’, 0 presented ‘<250’. Then Logistic Regression Model was used to fit the regressors and predict the probability of 1.

# 2. Data Exploration

## 1) Initial Exploration

All of variables were numeric data. The data was plotted as following: 

Through business analysis and instruction from question, there was no categorical property in those data. So, whole independent variables were suitable to fit linear regression model.

## 2) Missing Data

Although in Original Dataset, there was no obvious missing value, we found ‘-’ values in ‘TotRevSpend’ and ‘FanSatisfaction’ columns.

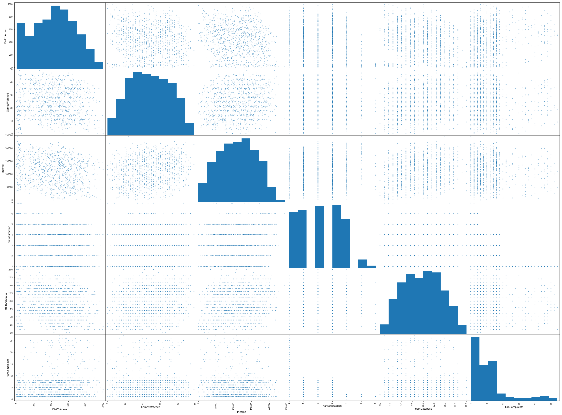
Because ‘TotRevSpend’ was the dependent variable, the instances where ‘TotRevSpend’ had missing values were impossible to be used to fit regression model. So, 40 instances where ‘TotRevSpend’ was 0 are deleted.

Through checking the correlation matrix of dataset (please see ‘3) Correlation Matrix’ ), the correlation between ‘FanSatisfaction’ and dependent variable was not very strong. Moreover, amount of remaining instances where ‘FanSatisfaction’ was 0 was 30, which counted for only 3.125% of whole dataset, we could ignore effects of those instances in modeling. 30 instances were deleted.

In sum, we cleaned all missing values in dataset, 70 instances counting for 7% of all data.

## 3) Correlation Matrix

The Correlation of Independent variables were shown in Scatter Diagram：



There was no obvious linear correlation between variables. We could see all of them as independent in-between.

Additionally，we checked the correlation between dependent variable and independent variables：

TotRevSpend 1.000000

YrsInDatabase 0.435460

GamesWatched 0.390067

Income 0.371446

FanSatisfaction 0.178089

FanComplaints -0.135933

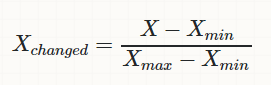
DistToArena -0.551067

We could see ‘YrsInDatabase’，‘GamesWatched’，‘Income’，‘FanSatisfaction’ had positive linear effects to TotRevSpend, but ‘DistToArena’ ,’ FanComplaints’ had negative effects.

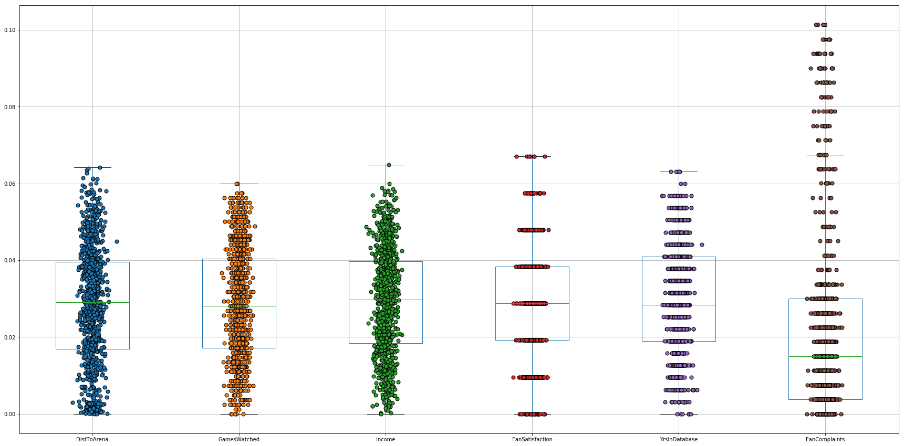
4) Outlier Detection

To check the Outlier, we plotted the independent variabls in boxplots.

Before plotting, we normalized all independent variables because the scale of variables varied largely.

Normalization Format: 

The boxplot was shown below:



Only data in ‘FanComplaints ’ had outliers outside of boxplot. However, the ‘FanComplaints’ variable recorded the complaint times a customer logged. Most of customers had small amount of complaints. So the high numbers of complaints were labeled outside of 75% of mean of data. Based on business understanding, dataset would lose essential information if those data point were deleted as outliers.

# 3. Linear Regression Modeling

To predict the ‘TotRevSpend’ based on all independent variables, we adopted several linear models to fit the data: Linear Regression, Ridge Regressoin, Lasso Regression. All Dataset was split into training set and validation set. The ratio between two parts was: training : validation = 80% : 20%

During modeling, we take advantage of Grid Search function to optimize the alpha in Ridge and Lasso Regression. The results of Model were listed as below:

Regression formula using Linear Regression:

TotRevSpend=294.383792577-0.500919306\*DistTArena+6.47784980\*GamesWatched+0.00100234636\*Income+7.08714975\*FanSatisfaction+16.9670242\*YrsInDatabase-1.04785478\*FanComplaints

Regression formula using Lasso Regression:

TotRevSpend= 295.42830646-0.500210192\*DistTArena+6.45811714\*GamesWatched+0.000998635599\*Income+7.00592279\*FanSatisfaction+16.9237078\*YrsInDatabase-1.01449921\*FanComplaints

Regression formula using Ridge Regression:

TotRevSpend= 294.729565341-0.494451651\*DistTArena+6.41039514\*GamesWatched+0.000997210169\*Income+7.22045609\*FanSatisfaction+16.8231530\*YrsInDatabase-1.05970416\*FanComplaints

To evaluate the performance of models, we calculated the Root Mean Squared Error (RMSE) and R-Square:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R-Square | | RMSE | |
| Training | Validation | Training | Validation |
| LinearRegression | 0.536429 | 0.514867 | 191.9284 | 195.3927 |
| LassoRegression | 0.536424 | 0.514869 | 191.9293 | 195.3923 |
| RidgeRegression | 0.536372 | 0.514942 | 191.9402 | 195.3776 |

The performances of three modeling were close to each other but not good enough. Additionally, the Performances of training dataset were better than that of validation dataset, which meant three models suffer from overfitting problem.

# 4. Logistic Regression Modeling

In order to predict probability that value in ‘TotRevSpend’ was bigger than 250, we created binary dependent variable and fit the data to logistic regression model.

The new binary dependent variable has 728 ‘1’s and 202 ‘0’s. The Ratio of ‘1’ to ‘0’ was close to 3.6.

In initial modeling, we keep 1-0 ratio when we split data into training and validation dataset. The ratio between two parts was: training : validation = 80% : 20%.Same as process in Linear Regression, we utilized Grid Search function to select best parameter of Logistic Regression.

Regression formula using Logistic Regression:

Z=0.00407168-0.00431588682\*DistTArena+0.0687240553\*GamesWatched+0.000010476\*Income+0.0221754336\*FanSatisfaction+0.122052739\*YrsInDatabase-0.0263899031\*FanComplaints

To evaluate the performance of Modeling, we produced confusion matrix for training and validation dataset:

Training Dataset:

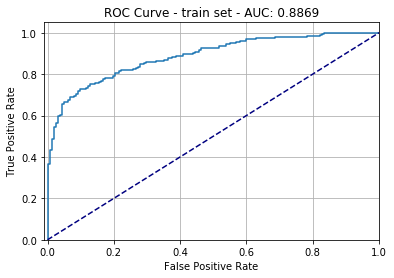
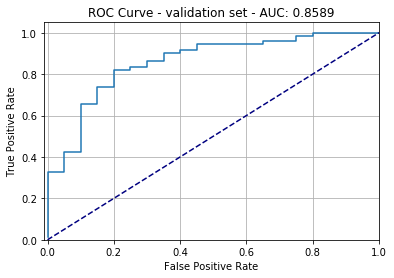
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | True Value | | Accuracy |
| 1 | 0 |
| Prediction | 1 | 543 | 84 | 86.60% |
| 0 | 39 | 78 | 66.67% |

Validation Dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | True Value | | Accuracy |
| 1 | 0 |
| Prediction | 1 | 67 | 9 | 88.16% |
| 0 | 6 | 11 | 64.71% |

We can see overall accuracy of model was not bad. Accuracy of ‘1’ was better than Accuracy of ‘0’ because the dataset was imbalanced. So the model intends to predict ‘1’ based on independent variables.

However, there were multiple ways to calculate the accuracy of binary prediction, which should be decided by business setting and business sense. To avoid misunderstanding, we adopt ROC curve to calculate AUC score to gain accuracy of model:

AUC calculated the area under ROC curve. The performance of model in validation dataset was actually a little worse than it in training dataset, which means the modeling suffered overfitting problem.

# 5. Further Investigation

Because of limitation of data volume, it is hard to ensure the accuracy of modeling using in practical problem. Not only Overfitting, metrics used to evaluate performance of modeling also show that data was not fitted well.

The most efficient way to improve the performance of models is to import mode data. The imported data should not only contain more instances of data, but also more features or independent variables. Extension of data volume could give model more labeled instances to fit the data. However, quality of imported data plays essential role. How to collect real and accurate data to supplement the current dataset.

Besides importing more data, regression modeling has different ways to avoid overfitting problem.

1) Linear Regression

To solve the overfitting problem, we could reduce the independent variables and check the metrics of modeling. In this specific problem, we removed the ‘FanSatisfaction’ and ‘FanComplaints’ because their low correlation with dependent variable: ‘TotRevSpend’. In other projects, p value to show significance of coefficients should be criteria to remove the dependent variables. However, Through analysis on significance of coefficients, all of feature are important to fit the model. That is the reason we remove two variables whose correlation with dependent variables and coefficients are relatively low.

After remove the two independent variables, the metrics of model are shown in following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R-Sqaure | | RMSE | |
| Training | Validation | Training | Validation |
| LinearRegression | 0.533654 | 0.512319 | 192.5019 | 195.9052 |
| LassoRegression | 0.533650 | 0.512309 | 192.5026 | 195.9079 |
| RidgeRegression | 0.533596 | 0.512408 | 191.5139 | 195.8874 |

Unfortunately, the performance of model become worse in both training and validation dataset, which should be normal because we remove significant feature, but the overfitting problem is still there. So, in current modeling, removing feature will not be useful.

We also could increase k in k-folder cross-validation to mitigate the overfitting. Now we setting the cross-validation attribute ‘cv’ in Lasso and Ridge regression from 5 to 50 to check the result:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R-Sqaure | | RMSE | |
| Training | Validation | Training | Validation |
| LassoRegression | 0.536068 | 0.514555 | 192.0030 | 195.4556 |
| RidgeRegression | 0.535298 | 0.514345 | 192.1623 | 195.4979 |

Unfortunately, the performance of model changed a little and overfitting problem was not solved.

Last but not least, we changed the modeling algorithm to check the result. We use XGBoost model, which was prevalent in prediction field, to train the dataset. The Result of XGBoost model was shown as follows:

RMSE: training dataset: 161.719, validation dataset: 187.854

We can see, although the metric improved, the overfitting problem was exacerbated. This’s mainly because the complexity of algorithm increased. Based on the worsen overfitting issue, the result from XGBoost could not be used.

In total, the best way to solve the overfitting problem in current dataset is augment of data volume.

2) Logistic Regression

Besides augment of data volumes, we try to adopt several techniques to made training set balance.

(1) Over-sample

One method to made imbalanced 1-0 label balanced was to over-sample the minority of binary dependent variable. In current dataset, we got 728 instances labeled as ‘1’ and 202 instances labeled as ‘2’. In those project, we select from 202 ‘0’ instances randomly to extend ‘0’ label to the amount same as ‘1’ label.

So, the final modified dataset has 1456 instances with 728 ‘1’s and 728 ‘0’s.

We split the modified dataset by 80%:20% and keep the ratio of ‘1’ : ‘0’ as 1:1 and then use training set to train logistic regression model.

The confusion matrix is shown below:

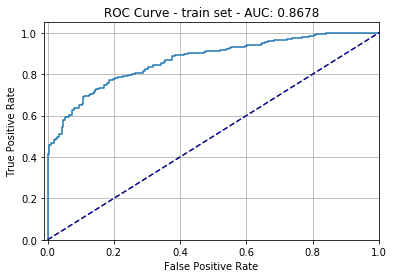
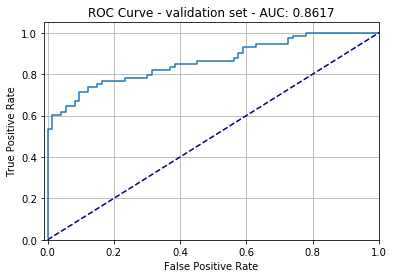
Training Dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | True Value | | Accuracy |
| 1 | 0 |
| Prediction | 1 | 449 | 109 | 80.47% |
| 0 | 133 | 473 | 78.05% |

Validation Dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | True Value | | Accuracy |
| 1 | 0 |
| Prediction | 1 | 54 | 10 | 84.38% |
| 0 | 19 | 63 | 76.83% |

The AUC score and ROC curve are shown below:

We can see the accuracy of 0 increases tremendous, but accuracy of 1 decreases a little. The result becomes more balanced and trustworthy. Overfitting Issue has been mitigated.

2) Under-sample.

On the contrary to Over-sample, we random selected instances from the dataset labeled as ‘1’ and matched the amount of instances labeled as ‘0’. Following this idea, we random selected 202 data from the ‘1’ instances and matched them with 202 ‘0’ instances as one group. Totally, we recreated 5 datasets in which the 1-0 labels were balanced. In modeling, we trained 2 level models.

To perform modeling, we used 5 different datasets to train 5 models separately in fist ensemble level including Linear SVC, Passive Aggressive Classifier, Logistic Regression, SGD Classifier and Random Forest Classifier. Then we combined the prediction of original dataset by those five models with independent variables in original dataset. Actually, the reason we conducted this modification was to introduce rectified features into dataset to correct imbalance issue. Then we keep ratio of 1-0 label to split combined dataset into training and validation dataset as stratification. Then we train second level logistic regression model to produce final predictions.

The results was shown in confusion matrix and AUC score:

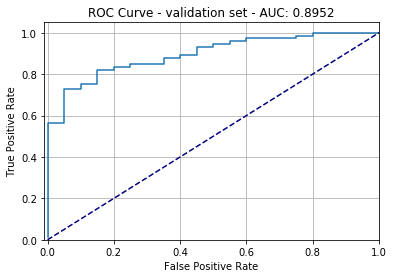
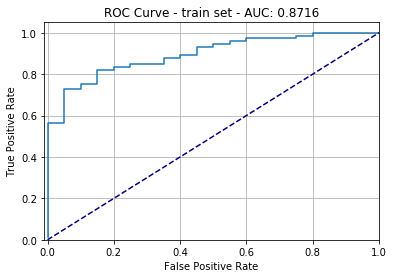
Training Dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | True Value | | Accuracy |
| 1 | 0 |
| Prediction | 1 | 539 | 90 | 85.69% |
| 0 | 43 | 72 | 62.61% |

Validation Dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | True Value | | Accuracy |
| 1 | 0 |
| Prediction | 1 | 70 | 11 | 86.42% |
| 0 | 3 | 9 | 75% |

The AUC score and ROC curve are shown below:



We can see, the ‘0’ prediction accuracy was similar with logistic regression without under-sample, but the AUC score was improved. Especially, overfitting issue was mitigated. The Total performance of 2-level ensembled model was improved.

Through investigation on current dataset, augment of data volume seems to be the best way to solve the overfitting problem in linear regression. In logistic regression, under-sample and over-sample could produce ideal result.

# 6. Model Selection

Through investigate of all regression models to predict numeric value in ‘TotRevSpend’, Ridge regression seemed to be relatively better model because it had lowest RMSE and similar R-Square comparing to other models.

Ensemble Model using under-sample method was chosen to predict probability of ‘TotRevSpend’ beyond 250 because it didn’t suffer from overfitting and evaluation metrics were competitive.