MONETIZATION OF ARTIST TRACKS USING PANDORA AND TWITTER DATA

Project By,
Rahul Raghavendiran (rahulrar@usc.edu)
EXECUTIVE SUMMARY

Objective:
To increase the profit of the musical company above the base profit by using insights from Decision Trees, Logistic Regression models, thereby finding attributes of the high propensity buyers from the 24,517 user data. The company is releasing the latest album from its artist and wanted to predict the potential buyer. They have planned to send the albums to 5,00,000 users in their database. The company's performance measure is to increase the profit above the base model and reduce the misclassification rate using the user information from Twitter and Pandora.

JMP Model:
In this model, we have extracted the features from the user data and have come up with the following variables like Gender, Age, Work Status, Region, Love Music, Listen Music, Artist Info, Own Artist, Artist Rating. Using JMP model, we have predicted the most influential variables that identifies the potential buyers. After initial analysis and survey of the customers the company has come up with the success rate of the users based on their previous album sales. It comes around 26.8% of their users bought the album. The company has used Logistic regression as the JMP model giving the base profit of $73,89,394.

Key Insights:
Overall the JMP model has a very good predictive power of 0.184%. The variables such as Age, Artist rating, Work status, Own Artist, Love music are the most influential variables in finding the potential buyers. The Summary Statistics of these variables provides very good insights for predictions. In order to predict the potential buyers from the training data, we used the scatter matrix plot, regression analysis for understanding the correlation and predictive power of these variables.

Proposed Model:
In our model, we have enriched the dataset with the additional variables which have huge impact in finding the potential buyers. The variables such as the Twitter follower of the artist, Number of retweets, like, share of the artist tweets, most influential genre of the user based on the Pandora data. We have also used the insights from the previous models and combined variables for better predictions. The misclassification rate is highly reduced and the predictive power of the has increased hugely by this strategy, and thereby we have accomplished our target of achieving profits above the JMP model.

Key Changes:
1) Data Enrichment by adding 3 probably influential variable which have high predictive power.
2) Combination of the variables, from the correlation info of the variables and by the knowledge from Decision tree and Logistic regression model.

Better Model:
We used the data enrichment strategy and the combination of the variables to built a logistic regression model thereby increasing the accuracy of the model for both the train and the test data. The misclassification rate is highly reduced, thereby increasing the predicted profit of the company. The Predictive power of our model is high compared and the variables are statistically significant with high chi square values. The accuracy of the predictions for the potential buyer has increased comparatively.
Lift Information:

<table>
<thead>
<tr>
<th>Lift Table in Dollars</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lift with respect to Baseline - JMP Model</td>
<td>4.77865527</td>
<td>4.1295537831</td>
</tr>
<tr>
<td>Lift with respect to Baseline - My Best Model</td>
<td>5.64024452</td>
<td>5.4893035469</td>
</tr>
<tr>
<td>Lift with respect to JMP Model - My Contribution</td>
<td>1.18029952</td>
<td>1.1487130231</td>
</tr>
<tr>
<td>Overall Lift with respect to Baseline - My Best Model</td>
<td>5.64024452</td>
<td>5.4893035469</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lift Table in Propensity</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lift with respect to Baseline - JMP Model</td>
<td>2.44403837</td>
<td>2.1959798995</td>
</tr>
<tr>
<td>Lift with respect to Baseline - My Best Model</td>
<td>2.773300046</td>
<td>2.7156173617</td>
</tr>
</tbody>
</table>

Conclusion:

Our model has a very high R-square value which implies the better predictive power of our model. The misclassification rate for both the test and train data of our model is low compared to JMP model. This increases the accuracy of the model in predicting the high propensity buyers. We have used the insights from both decision trees and Logistic regression for coming up with the high predictive variables and have achieved the goal of increasing the profit of the company by 20%. In our model, all the variable are statistically significant and also have high chi-square value. The ROC curve of our model covers more area than the JMP model which shows high accuracy in the prediction of the potential buyer. Therefore, our model better compared to the JMP model in various aspects and achieved the target of beating the base profit of JMP model.
TABLE OF CONTENTS

SECTION 1.............................................................. INTRODUCTION
SECTION 2.............................................................. PREPROCESSING RAW DATA
SECTION 3.............................................................. DATA
SECTION 4.............................................................. ETL PROCESS
SECTION 5.............................................................. ANALYSIS
SECTION 6.............................................................. PERFORMANCE MEASURE
SECTION 7.............................................................. BUSINESS INSIGHTS
SECTION 8.............................................................. BIBLIOGRAPHY
INTRODUCTION

The Musical company which owns the license to sell the artist's album is looking to make use of the social media data of their customers in their database so that they can predict the potential buyers for their upcoming album release. The musical company has more than 1 Million users in their database. They have decided to send the new album package to 5,000,000 users in their database. It is known that based on their previous album release they have the data of who were the buyers of the album. Out of 24,517 users whose Pandora and Twitter data are available, the potential buyers were only 26.7%. The company wanted to improve their base profits using data mining techniques. The dataset that the company have, contains Twitter information such the tweet information of the users about the artist, Age, Region, Gender, Work status of the user. The company also have got information from the Pandora radio station, which includes information such as how much the users loves listening music, how much hours the user listen to music daily, how much information the user knows about the artist, how many music songs the user owns in his/her playlist. Now for the initial analysis, the company has hired a data scientist for designing a model to increase the base profit for their new album release. The company projects the following pricing,

- Selling price of the album = $40
- Cost price of the album = $10
- Salvage rate per package = $10
- Cost of mailing the product = $4.50
- Cost of returning the product = $4.50
- Profit per product sold = $25.50
- Loss per product sold = $5.00

PREPROCESSING RAW DATA

As a data scientist, we have to come up with the dataset to analyze from the raw data.

- In order to decode the tweets of the user about the artist, we did sentiment analysis for predicting the sentiment of the user towards the artist. Based on the intensity of the tweets compared to all others, we have come up with the artist's rating of the user. This information is used as one of the factor in predicting the Artist Rating variable.
- **Work Status** information, from the basic info from Twitter which is in the form of a comment, we have decoded these comments into 3 categories such as Employed, Student, Others (retired, unemployed, social service). We have used few Natural language processing (NLP) technique to decode these comments of the users.
- For the **Region** information, we had the state information, we recoded to region wise for better performance of the variable. After recoding the predictive power of the variable increased. We recoded the variable to 3 regions North, South, Centre.
- For **Music Love** information, we were given the opinion of the users about the role of music in their day today life. We decoded the comments using the NLP technique to come up with the categories such as little, moderate, great.
- For **Artist Info** variable, we were given the users comments on the artist. We used NLP techniques to decode and categorized into 'more', 'less'.

• For **Own artist** tracks, we were given the number of playlist songs and bookmarked songs the users had in their integrated Pandora account. We need to come up with a classification model to split these values into 4 categories such as High, Moderate, Little, None.

• For **Artist rating**, we used values such as the number of songs the user listens to the artist's song per day, how many bookmarked songs, rating based on the tweets, active user of both Pandora, Twitter, Love for music. All these variables are used to come up with the rating in the scale of 0 - 100. More information is explained in the next two sections.

**DATA**

The most promising variables that determine the high propensity buyers are the following,

**Artist Rating:**

![Summary Statistics for Artist Rating](image)

This variable is a **quantitative**. From the summary statistics, it is clear that the average rate of liking from the user is 47.5 which implies the artist's tracks ratings are around 50. The scatter matrix plot above shows the distribution of the variable with respect to success, which gives us the insight that the users who have rated from 40 to 50 and 70 to 80 have the high propensity buyers.

**Age:**

![Summary Statistics for Age](image)

This variable is a **quantitative**. From the summary statistics, we come to know that the average age of the train data is around 37 and from the scatter matrix plot, it is clear that the potential buyers are uniformly distributed across from the age of 13 to 70. Therefore this variable doesn't provide any insight from these above informations.

**Own Artist:**

![Frequencies for Own Artist](image)
This variable is a **qualitative**. From the summary statistics, we could see that most of the users are not aware of the artist and have no tracks in their playlist. This variable is recoded to 4 categories based on the Pandora user information. The categories are none, little, moderate, high number of tracks of the user. From the scatter matrix plot, it is evident that the potential buyers are mostly in little and moderate categories compared to others and the interesting insight is that the users who haven't listen to the artist tracks show interest in buying the album.

**Artist Info:**

![Frequencies](image1)

This variable is a **qualitative**. From the summary statistics, we don't find any business insights apart from the fact that most users had enough info about the artist. From the scatter matrix plot, we find an interesting fact that the users who had lesser info about the artist is more probable to buy the album.

**Work Status:**

![Frequencies](image2)

This variable is a **qualitative**. From the summary statistics, we find that around 60% of users are employed and the next best category is the students with around 16% of the users. From the scatter matrix plot, it is evident with the insight that employed users have high propensity to buy. The next best category to concentrate is that of the students.

**ETL – EXTRACT TRANSFORM LOAD ENRICH**

Initially, we came up with the data using the Twitter and pandora APIs. The features extracted are explained briefly in the below table,

<table>
<thead>
<tr>
<th>USERID</th>
<th>User ID to identify the unique user and their artifacts.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENDER</td>
<td>Gender info about the user. This information is obtained from the Pandora user profile information. This column has been recoded into 1(Male) and 0(Female)</td>
</tr>
<tr>
<td>AGE</td>
<td>Age info of the user. This information is obtained from the Pandora user profile information.</td>
</tr>
<tr>
<td>WORK STATUS</td>
<td>Work status of the user. This information is obtained from twitter and Pandora account.</td>
</tr>
</tbody>
</table>
This column has been recoded into 3 categories after the analysis of the work status data. The categories are employed, student, others.

**REGION**

(qualitative) Location of the user. This information is obtained from the pandora user profile information. This column has been recoded into 3 categories after analysis of the location data. The categories are North, South, Center

**LOVE TO MUSIC**

(qualitative) The user's passion for music. This information is newly created for the dataset based on sentiment analysis of the user's tweets about the artist tracks and the information of the user's pandora profile such as... number of hours of listening to artist songs etc... This column has been recoded into 3 categories like Great, Little, Moderate

**LISTEN_MUSIC**

(quantitative) Average number of hours the user listens to music in Pandora daily. This information is obtained from the Pandora user profile information.

**ARTIST INFO**

(qualitative) How much info the user knows about the artist. This information is obtained from both the twitter and Pandora info. This column is newly created from the sentiment analysis of the user's tweets and the information of the user's pandora profile. This column has been recoded into 2 categories like more, less.

**OWN_ARTIST**

Average number of hours the user is listening to the artist music in pandora. This is information is obtained from Pandora profile and recoded to categories like none, little, moderate, high.

**ARTIST RATING**

(quantitative) Artist's Rating of the user after listening to tracks. This information is newly created for the dataset based on sentiment analysis of the user's tweets about the artist tracks, number of retweets, shares and liking and the information of the user's pandora profile such as number of bookmarks, etc...

**SUCCESS**

(quantitative) This tells the buyers and non buyers of the previous album sales. This information is obtained from the musical company database. It is either 1(buyer) or 0(non buyer)

We have more than 1 million tweets about the artist and the user information from Twitter and Pandora. We have done intense analysis on how to decode these tweets and what features to extract from the datasets. Then we finalized with features which are mentioned in the above table. Most of the features are obtained from the Twitter, Pandora integration of the users. Most of the variables are recoded, indicated in the table above, from the raw data extracted by a script which used the Twitter and Pandora APIs. The recoding is done based on the comments of the users in the Pandora profile and from the twitter information of the user. After complete analysis of the dataset, we came up with the 15,000 tuples as train data and 9517 tuples as test data.

**Feature Extraction**

**Work Status:**

This variable is extracted from the background information of the user from Pandora. The raw data contained users comment on their work status. We have come up with the categories based on the comments.

**Region:**

The raw data just had the states of america, however we did an initial analysis of the variable and found the insight that categorizing the variable as North, South, Centre plays a key role to find the high propensity
buyers, instead of North, South, East, West.

**Love To Music:**

This information is extracted from both the Pandora and Twitter info about the user. Based on the Twitter info of the user such as number of the following of the music artist, no. of artist retweets, share, likes and from the pandora information such active usage of the account, total number of hours of listening to music.

**Artist Info:**

This column is newly created based on the information from both Twitter and Pandora. The Artist info of the user is obtained from the tweets of the user about the artist and the number of days the user had listened to the artist tracks from Pandora. After this initial analysis, we recoded the information into 2 categorize such as known more or Known less.

**Artist Rating:**

This column was newly created based on the tweet information of the user about the tracks of the artist. We have done Sentiment analysis of the tweets and have also used few other informations like number of retweets, share, like from Twitter. In addition to determine this variable, we also made use of the Pandora information such as number of bookmarks of the artist tracks, number of songs of the artist in the playlist.

**Own Artist:**

This column was newly created from the Pandora user profile. This column is obtained based on the number of the bookmarked artist songs, number of artist songs the user has in his/her playlist. Based on these info, we have recoded the values to categories such as little, moderate, high, none.

**ANALYSIS**

In order to come up with a better model than the JMP model, we have used the following techniques,

- Correlation information of the variables,
- linear regression to fit the variables,
- Scatter Matrix plot,
- Decision trees for the entropy information of the variables,
- Logistic regression for finding the most influential and highly predictive variables,
- Enrichment of the users data
- Combination of the users attributes recoded as a new variable.

**Enrichment of the users data :**

We have used the Twitter and the Pandora API for enriching the current data set. The enriched variables are Twitter Follower, Twitter Share and probably most influential Genre of the user information.

**Twitter Follower:**

![Frequencies](image1)

![Scatter Matrix](image2)
This variable is quantitative. From the summary statistics, we see that most of the user are Twitter follower of the artist. This variable has high predictive power for the potential buyers since the scatter plot shows that most of the buyers of the album follows the artist. The regression analysis also showed the predictive power of the variable to be high.

Twitter Share:

![Twitter Share Frequencies](image)

This variable is qualitative. The summary statistics shows that most users moderately retweet, share and like the artist tweets. From the scatter matrix plot, it is interesting to find that users who shared moderately were the potential buyers.

Genre:

![Genre Frequencies](image)

This variable is qualitative. The summary statistics shows that most of the user's preferred genre is the Blue. The next best genre is that of Jazz. From the scatter plot, it is evident that the users who liked 'blue' were the potential buyers. It is interesting to see that the 'others' category to be empty. The regression analysis showed the predictive power of the variable is high.

Combination of variables:

- After an intense analysis of the variables, decision tree model and Logistic regression, we find the following insights. The analysis from the decision tree model provided info that the combination of Artist Rating and Age have a high predictive power for predicting the potential buyers and their scatter plot also provided with very good information. This variable is quantitative. This variable is statistically significant.
- The other combination is that of Work status and Own Artist track variable. The regression analysis of this provides us with the high predictive information of the variable. This variable is quantitative. This variable is statistically significant.
Best Model for the Enriched Data set:

Profiler:

From the profiler, the enriched variables such as Twitter Follower, Genre plays a key role in predicting the potential buyers. The combined variables such as WORK_OWN also predicts the potential buyers.

Statistical KPI:

The predictive power of the model is around 40% very high compared the JMP model. The variables used on this model are all statistically significant and enriched variables have have high chi-square values, which shows their high predictive power. The misclassification rate of this model is low compared to the JMP model. Therefore the accuracy of the model is quite high in predicting the potential buyers.

ROC and Lift Curve:

The ROC characteristics of the model shows that area under the curve is high compared to the JMP model which predicts a high profit and less misclassification rate. The Lift curve provides the lift of our model.

When considered top 50% of the users then our model provides a lift of 2 times compared to the JMP model.
PERFORMANCE MEASURE

To measure the performance of our model we use the misclassification rate and the profit of our model as the performance measure. We have shown the snapshot of the measure for both the training and the testing data.

**Training:**

The Accuracy of our model is comparatively high to the JMP model. The profit we obtained is 1.2 times higher for our model compared to the JMP model. We also have achieved the target of shipping 5,00,000 packages to the customer. The False Positive rate is very low compared to the JMP model. We have also increased the propensity to buy the album to 73.5% which is 1.12 times than the JMP model. The True positive rate of our model has increased 100% than the JMP model.

**Testing:**

The Accuracy of our model is comparatively high to the JMP model. The profit we obtained is 1.2 times higher for our model compared to the JMP model. We also have achieved the target of shipping 5,00,000 packages to the customer. The False Positive rate is very low compared to the JMP model. We have also increased the propensity to buy the album to 73.5% which is 1.12 times than the JMP model. The True positive rate of our model has increased 100% than the JMP model.
The Accuracy of the validation set is also comparatively high than JMP model. The values in the above image shows that our model has very low misclassification rate and the model din't over-fit the training data. We also increased the propensity of the buyers like 1.23 times higher. We also have achieved the target shipping of 5,00,000 packages to the customer. The True positive rate has increased 1.5 times than the JMP validation set. It is evident from the above figure that the false positive rate is very low for our model.

**BUSINESS INSIGHTS**

- The Correlation between the variables and the regression analysis is used to predict which combination of the variables work and the high R-square value suggests the predictive power of the variable. For example, the variables like Twitter Followers, Genre, Artist Rating, Work_own, Own Artist Tracks are presented with high R-square values which provides us the high propensity buyers.
- The most influential variable is 'Twitter Follower', which is statistically significant. It has very high predictive power for finding the potential buyer.
- The 'genre' of the users is the second best attribute, it also has high Rsquare value and the chi-square value. The scatter plot of the variable shows us that the users whose genre is same as the artist i.e blue have high propensity to buy the album.
- The combined variable 'Work_OWN' using the decision tree and Logistic regression models have high predictive power and we see that from the scatter plot the employed works who moderately own the songs of the artist have high propensity to buy the album.
- The variable 'Artist Rating' also have high predictive power and high chi-square value. From the scatter plot it is evident that the user who has rated the artist in the range 40 to 50 and 70 to 80 has the high propensity to buy the album.

**BIBLIOGRAPHY**

- [www.Pandora.com](http://www.Pandora.com)
- [www.Twitter.com](http://www.Twitter.com)
- [www.wikipedia.com](http://www.wikipedia.com)
- [www.jmp.com](http://www.jmp.com)