ANALYSES OF TICKET RENEWAL IN NFL

Anirudh Tunga, Dheeraj Peddireddy, Gokulakrishnan Swaminathan, Shubham Arora, Matthew A. Lanham
Purdue University, West Lafayette, IN 47907
atunga@purdue.edu; dpeddire@purdue.edu; gswamina@purdue.edu; arora90@purdue.edu,
lanhamm@purdue.edu

ABSTRACT

In the past few years, sports analytics has penetrated many sports including NBA, NFL, NHL and data driven decisions have become an integral part of most of the sports. In this project, we are analyzing the ticket renewal data provided by the client and aid them to identity the key metrics that are leading to the renewal of their season ticketholders year over year. Furthermore, our approach will help them to target specify segment of fans to target in future based on the identified metrics leading to their renewal. Overall, we will adjudicate the impact of the new renewal benefits and timeline provided them.

We are building a predictive analytics solution combined with insights that can be directly generated based on the initial data itself. We will perform different techniques like Logistics Regression, Random Forest, SVM etc. and evaluate the performance of each of those models taking in ‘renewal’ as the dependent variables. We further plan to perform clustering to segment the fans and compute common features (behaviors, demographics etc.) to target each of those segments separately based on the computed common features. We are using R studio as our primary coding platform with the basic EDA and visualizations in Tableau.

INTRODUCTION

The amount of season tickets sold for the season can account for a major part of total paid attendance for any major league sports team. With this in mind a manager may wonder what brings season ticket buyers back season after season, and what can be done to measure this occurrence. An added question for front office staff members is, do these reasons coincide with a team’s marketing strategy to maximize the number of fans who buy season tickets?

The client has changed their entire ticketing renewal strategy for the upcoming 2019 season. We are trying to identify what is helping the client to renew their season ticketholders year over year, What segment of the season ticket holders were most receptive to the new timeline and incentives? We are trying to identify which segments of fans, or actions on the accounts are leading to more prevalent season ticket renewals. Leveraging the statistical and machine learning tools available, we are going to predict the factors which govern a customer to renew the season ticket and further explore the reasons behind the churn.

Our approach can be divided into two categories:

1. Forecasting algorithm development
2. Explanatory variable configuration

For algorithm development, the following three subgroups exist: (1) statistical learning models, including linear regression and probabilistic models; (2) time series forecasting models such as new product diffusion models and the vector auto-regression (VAR) method and (3) machine learning-based models such as artificial neural networks (ANN). Meanwhile, in the explanatory variable configuration, the main topic is the way in which sentiments can be incorporated into explanatory variables.

Improvement can be achieved by adopting sophisticated machine learning based nonlinear regression algorithms. Because most forecasting models are based on linear algorithms, they cannot determine nonlinear relationships that may increase the forecasting accuracy. Only a few trials have attempted to develop forecasting models based on nonlinear algorithms; The employment of various well-designed nonlinear regression algorithms, such as support vector regression (SVR), Gaussian process regression (GPR), and k-nearest neighbors (k-NN), and the combination of individual forecasting models, can enhance the forecasting accuracy.

The scope is not only about retaining the existing customers but also about bringing in New customers. For the new customers, they expect the organization to treat them with good seats and other additional privileges. So, for an existing stadium with the limited number of good seats, there's a trade-off in allocating new seats to the anticipated customers. Hence the model is built with a careful consideration with this idea as the growth of a sports organization depends on the growth of purchase of season ticket renewal.

Even though a prediction model is built and the areas of improvement are identified, the question arises to each and every individual is that how efficient models perform in areas where each and every variable is not consistent, correlated and time dependent. Does this mean that every time a model is to be updated for each time there is change in the features? If yes, what would be the accuracy and the frequency of change that must be accounted for?

Apart from this what would be effects of economic and social implications imposed on the organization to the proposed solutions. That is, though the results of model suggest improvements in the chosen features, how likely that would improve the current economic and social condition of the organization. Because both attributes are most important to for customer-service organization.

With these various questions in mind, we are proceeding to understand the behavior of the features and its implications to the renewal rate.

Reference:

Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes, Jack V. Tu

LITERATURE REVIEW

The below papers discuss the behavior of consumers and understanding the season ticket holder's perspective while renewing the tickets. This paves the way for the immense scope of analytics which can be done on tennis.

Prediction of ticket purchase in professional sport using data mining, Chen-Yueh Chen

The primary purpose of this study was to predict the types of tickets the fans would purchase to attend the home games. The model used for prediction was done using a multinomial probit model named CRISP. The data analysis is done using various parametric and non-parametric models like discriminant analysis, logistic regression, decision trees, artificial neural network, collaborative filtering, market-based analysis, survival analysis, genetic algorithms and so on. The model is trained and validated using the past data which is divided randomly based on the research design. The prediction parameters found from the results of the various models include the type of opponent faced, late season, promotion and the value of the team. The prediction accuracy of the model developed was 60% which was twice the value from the existing prediction model.

Industrial Marketing in Sport: Understanding Season Ticket Renewal Across Account Types Clinton J. Warren School of Kinesiology & Recreation, Illinois State University

The objective of this study is to advocate for a new conceptual foundation on which future research and practical season ticket sales and marketing initiatives can be built. It further segments season ticket holders based on the resulting Season Ticket Holder's account types and analyzes the differences across account types on critical variables affecting renewal intention. Season ticket holders are segmented using k-means clustering and further the group difference among these segments are conducted based on product quality, relationship quality, and perceived constraints to renewal. Finally, an overall intention will be analyzed based on all the segments. The study cites the distinction between customers, consumer & fans and how all of these terms have different meaning in sports industry. The make use of a one-way multivariate analysis of variance (MANOVA) using the STH groups from the k-means cluster analysis as the independent variable and product quality, relationship quality, and renewal constraints as dependent variables.

The Factors Influencing Churn Rates among Season Ticket Holders: An Empirical Analysis, Heath McDonald, Deakin University

This study tried to identify and measure the main factors posited to explain churn among Season Ticket Holder (STH), especially those relevant to a high-involvement product like sports. It aims to determine if attitudes and/or behavior can explain churn in sport organizations, as they have in traditional repertoire markets. It further finds the nature and strength of the between these factors and actual churn rates. The study has considered mostly the first year STH to investigate the importance of length of relationship to churn rates. It tries to gain insight into the development of loyalty & commitment by modeling the attitude and behavior of STHs across wide range of experiences behaviors of STHs across a wide range of accumulated experiences, we expect to gain insight into the development of loyalty and commitment among relational customers. In addition, by examining STHs with varying levels of direct contact with the organization (e.g., attendance levels), we hoped to distinguish between those who form bonds in an active way, and those who are more passive.

A User Behaviour-based Ticket Sales Prediction Using Data Mining Tools: An Empirical Study in an OTA Company, Chengfu Yang School of Information

This paper develops an integrated forecasting model by combining various internal and external factors that influence ticket sales and its market value. A feature selection model using machine learning algorithms is employed to get an accurate prediction of the user behavior for ticket purchase. The model described in the paper uses the concept of Neural networks and the Support Vector Regression. From the model, we can understand the two factors that influence the traditional OTA: the calling number of OTA and the ticket search query. Although these factors are only a small portion of the organization, it greatly influences the ticket sales. Also, the external factor information cannot be utilized completely to predict due to the boundedness of the sample.

A regression-based predictive model of student attendance at UVA men's basketball games (T.L.W. Walls ; E.J. Bass)

The objective of this study is to develop a regression-based predictive model to allow better prediction of attendance for the student general admission seats at University of Virginia men's home basketball games. The best six variable model contained the following factors: whether UVA was ranked, whether the opponent was ranked, opponent popularity, and whether classes were in session. This model resulted in an adjusted R^2 value of 0.816; for modeling attendance behavior. A one-sided Mann-Whitney test revealed that the model errors were significantly small. The use of regression-based approaches for predicting attendance is a promising method, especially for longer term planning when ticket sales data may not yet be available.

Measuring season ticket holder satisfaction: Rationale, scale development and longitudinal validation, (Heath McDonald, Adam J. Karg, Andrea Vocino)

Season tickets are examples of sports subscription products, research into which is very limited. Given the nature of subscription markets, there is sufficient reason to expect that the relationship between service quality, satisfaction & renewal might operate differently from transactional markets. This paper seeks to address this deficiency in the research by developing & verifying a scale that identifies the components of professional sport club season ticket packages that are most influential on buyer satisfaction. Survey data were collected over three consecutive years from season ticket holders (STHs) supporting the same team. The result is a 19-item scale measuring overall satisfaction as well as five key constructs by which STHs assess the season ticket package: service, home ground, on-field performance, club administration, and personal involvement.

Predicting season ticket holder loyalty using geographical information, (Dominik Schreyer, Sasha L. Schmidt, Benno Torgler)

This article addresses the notable shortcoming that sports economists have not explored the potential determinants of STH loyalty as expressed through regular stadium attendances. It does this by exploring the potential determinants of STH stadium attendance demand. Particularly, it analyses the less researched role of increasing opportunity costs resulting from larger home-stadium distances in STH stadium attendance demand. The results suggest that STHs' geographical location plays a crucial role in predicting STH stadium attendance demand. Moreover, a suprising relationship between nonlinear distance–attendance relationship, indicating that behaviourally loyal STHs live either exceptionally close or far away from the stadium.

The demand for NFL attendance: A rational addiction model study:

This paper examines the demand for attendance at NFL games using a rational addiction model to test the hypothesis that professional football displays the properties of a habit-forming good. It is found that past and future attendance, winning percentage, the age of the stadium are the significant factors in the determination of attendance at NFL games. During modeling, it was observed that the winning percentage in the home games greatly influenced the model. The results generated from the model are consistently confirming the hypothesis.

Summary of the literature review

Literature Review	*Context/Purpose*
Prediction of Ticket Purchase	Predict types of tickets fans would purchase to attend the home games
Industrial marketing in sports	Cites the distinction among customers, consumer & fans and how all these terms have different meaning in sports industry
The factors affecting Churn Rates	Identifies and measure the main factors posited to explain churn among Season Ticket Holder (STH)
A User-Behaviour based Ticket Sales	Develops a forecasting model by combining various internal and external factors that influence ticket sales
A regression based predictive model	Develop a predictive model for better prediction of attendance for the student general admission seats at University of Virginia men's home basketball games
Measuring season ticket holder satisfaction	Developing a scale that identifies the components of professional sport club season ticket packages that are most influential on buyer satisfaction
Predicting season ticket holder loyalty using Geographical Info	Suggest that STHs' geographical location plays an important role to predict STH stadium attendance demand
The demand for NFL attendance	Examines the demand for attendance at NFL games using a rational addiction model

DATA

The data comes from one of the NFL club and thus is proprietary in nature. Overall, it consists information related to Seasonal Ticket Holders spread across 3 different tables Renewal information, Survey Data and 3rd party data provided to the club. Table 1 consists of information like Tenure, Zip Codes of account holders, Base seats renewed , Date of renewal etc. The 3rd party data consist of demographic information like Gender, Age Groups Education, Gender etc. Lastly, the survey data has information related to questionnaire asked from the account holders and their responses. We have merged the Renewal and Turnkey data based on the AccountID (consists of 3948 unique account IDs)

Table 1: Renewal Data

Field	Type	Description
AcctId	numeric	Ticketing AccountID number
Tenure	numeric	Number of years being a Colt's season ticket member
AttendancePercentage	numeric	The 2018 attendance percentage for Colts regular season games
Zip	factor	Zip code of the account holder
CountyName	factor	County in which the account holder is from
State	factor	State in which the account holder is from
DistanceFromStadium	numeric	How far the account holder's zip code is from Lucas Oil Stadium
NumberOfCalls	numeric	Number of calls a customer service rep has made to this account during the 2019 renewal program
NumberOfVoicemails	numeric	Number of voicemails a customer service rep has made to this account during the 2019 renewal program
NumberOfEmails	numeric	Number of emails a customer service rep has sent to this account during the 2019 renewal program
BaseRevenue	numeric	Season ticket revenue that is up for renewal for the 2019 season (their 2018 season ticket account value)
BaseNumSeats	numeric	Season ticket number of seats that is up for renewal for the 2019 season (their 2018 season ticket number of seats)
BaseSection(s)	factor	The account holder's season ticket section(s) that are up for renewal for the 2019 season - comma separated could be multiple
BaseRow(s)	factor	The account holder's season ticket rows that are up for renewal for the 2019 season - comma separated could be multiple
BasePriceCode(s)	factor	The account holder's season ticket p[rice codes that are up for renewal for the 2019 season - comma separated could be multiple. A price code goes from D - O (D being the highest priced seats, O being the lowest priced seats)
RenewedRevenue	numeric	Season ticket revenue that has been officially renewed for the 2019 season
RenewedNumSeats	numeric	Season ticket number of seats that has been officially renewed for the 2019 season
RenewedSection(s)	factor	Season ticket section(s) of seats that has been officially renewed for the 2019 season - comma separated could be multiple
RenewedRow(s)	factor	Season ticket row(s) of seats that has been officially renewed for the 2019 season - comma separated could be multiple
RenewedPriceCode(s)	factor	Season ticket price code(s) that has been officially renewed for the 2019 season - comma separated could be multiple
RenewalDate	factor	Season ticket renewal data for the 2019 season

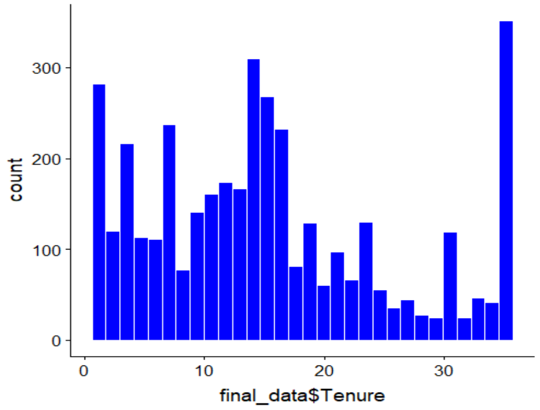
Table 2: Turnkey Data

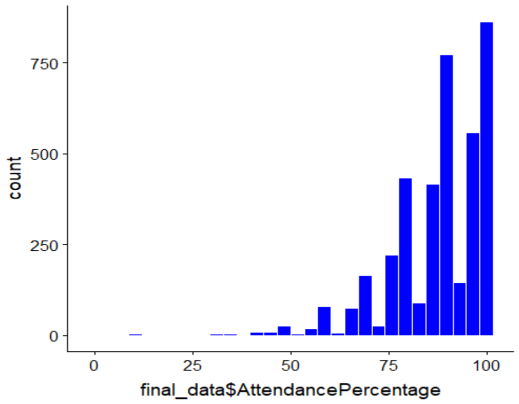
Field	Type	Description
TicketingSystemAccountID	numeric	Ticketing AccountID number (same one as the Renewal Dataset AcctId)
AgeIncrements	factor	Age Range of the account holder
Gender	factor	Gender of the account holder
Occupation	factor	Occupation the account holder currently partakes in
Education	factor	Education level of the account holder
MaritalStatus	factor	Marital status of the account holder
EstHhldIncome	factor	Estimated household income level of the account holder
HomeOwnerRenter	factor	Indicator for if the account holder owns or rents their place of residence
LengthofResidence	factor	How long the account holder has lived in their current residence
HomePropertyTypeDetail	factor	What type of property the account holder currently resides in
HomeMarketValue	factor	Market value of the account holder's current residence
Vehicle	factor	What type of vehicle the account holder currently owns
DiscretionaryIncomeIndex	factor	Indicator for how much discretionary income the account holder has (the higher the value the more discretionary income they have to spend on non-necessities) - Keep in mind our product is driven by this portion of income
HomeAssessedValueRanges	factor	What the account holder's home was assessed at
NetWorthGold	factor	The net worth of the individual

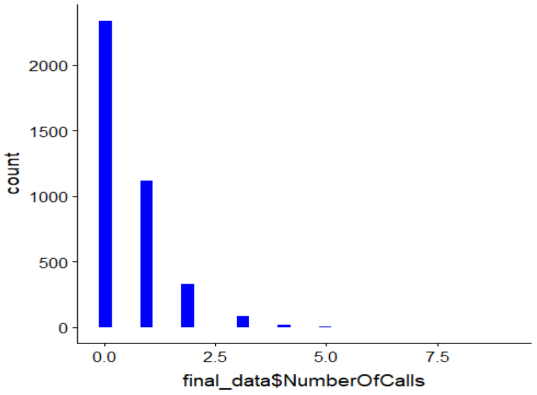
Table 3: Survey Data

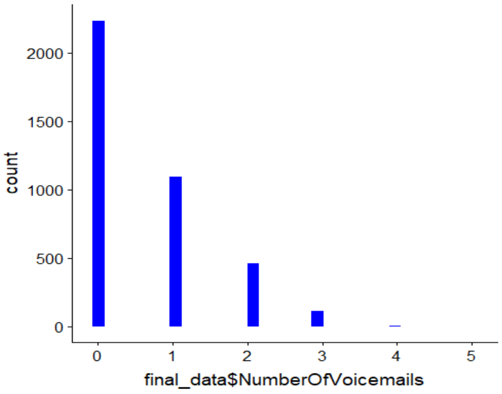
This table consists of data related to surveys conducted by client to get a feedback from the Seasonal ticket members regarding their satisfaction level, would they recommend this to someone else, does the club meet the expectations, if their problems are resolved etc.

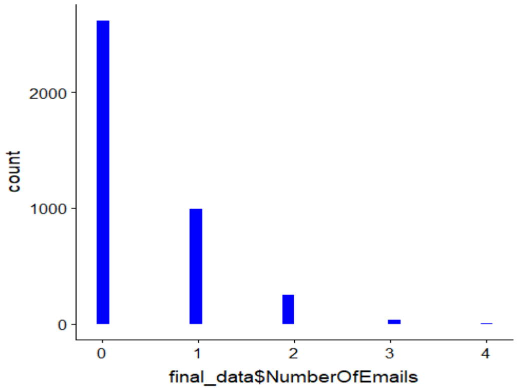
Univariate Analysis of different variables

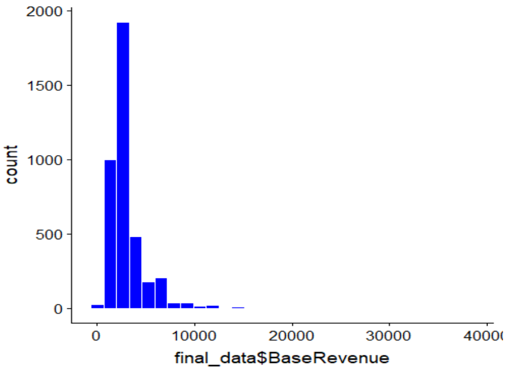


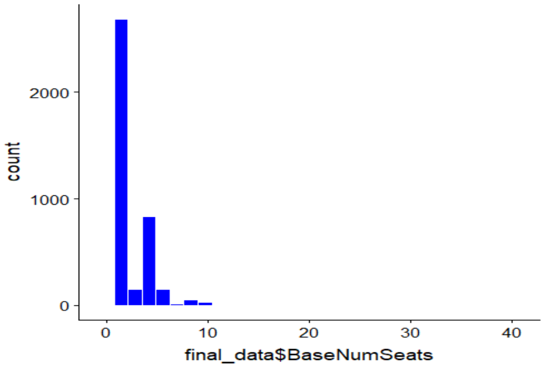


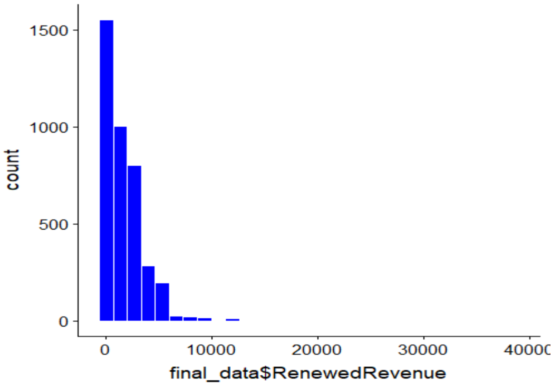

count
0
500
1000
1500
2000
0.0
2.5
5.0
7.5
final_data$NumberOfCalls


count
0
500
1000
1500
2000
0
1
2
3
4
5
final_data$NumberOfVoicemails


count
0
1000
2000
0
1
2
3
4
final_data$NumberOfEmails


count
0
500
1000
1500
2000
0
10000
20000
30000
4000(
final_data$BaseRevenue


count
0
1000
2000
0
10
20
30
40
final_data$BaseNumSeats


count
0
500
1000
1500
0
10000
20000
30000
4000C
final_data$RenewedRevenue

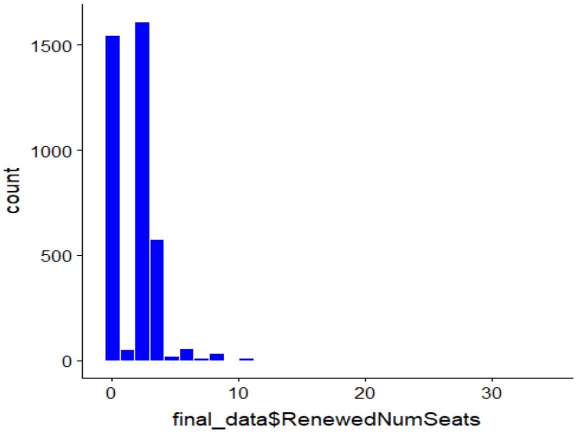

count
0
500
1000
1500
0
10
20
30
final_data$RenewedNumSeats

Figure 1

Renewed/Base Revenue based on number of emails/voicemails/calls

Based on Figure 2, all of these variables including number of emails/voicemails/calls have a negative relation with the renewed as well as the base revenue.

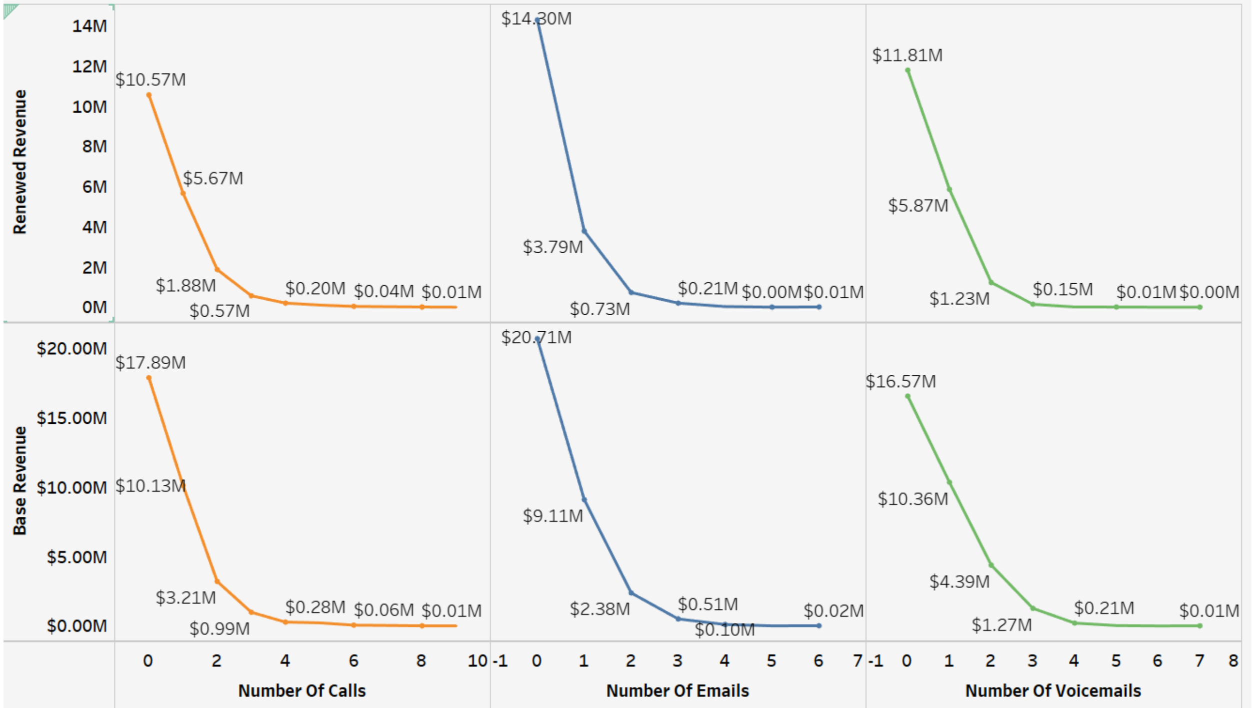


Figure 2

Heatmap for renewed Revenue for different county

The location plays an important role while driving the renewals as the revenues are more for the places in vicinity of stadiums. Figure 3 shows some of the highest revenue generating counties.

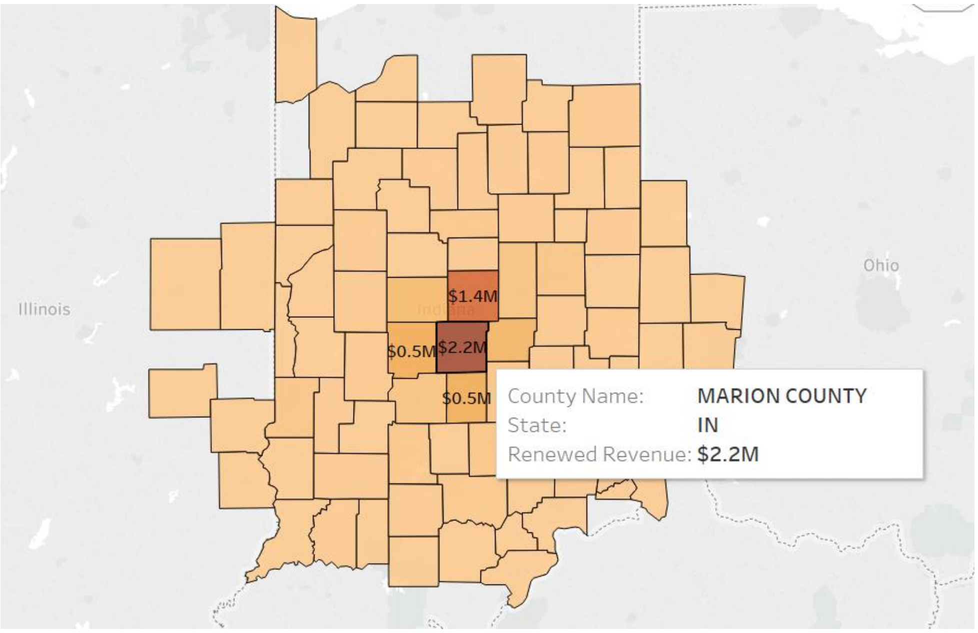


Figure 3

Month on Month Renewed Revenue (in Millions) Comparison

Most of the renewals have been done in the month of November compared to other months. There is an increasing trend which shows more renewals are done in the end of the month.

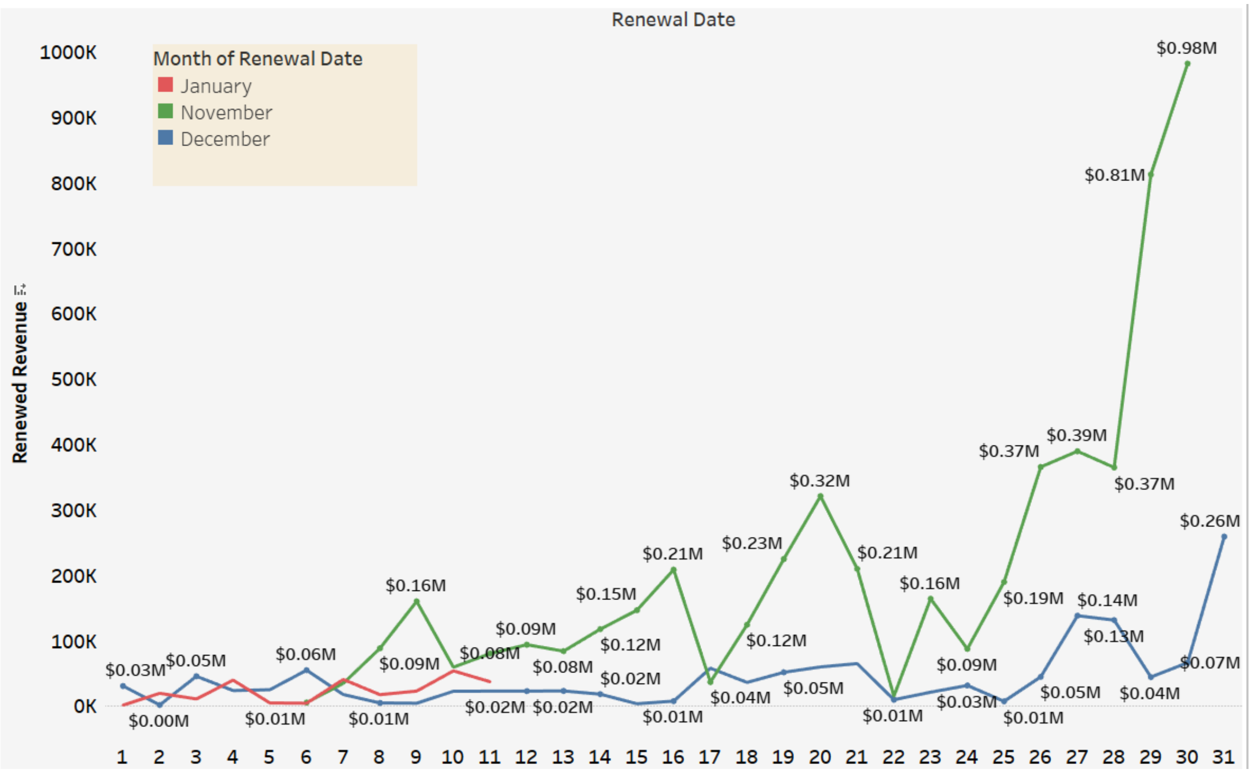


Figure 4

METHODOLOGY

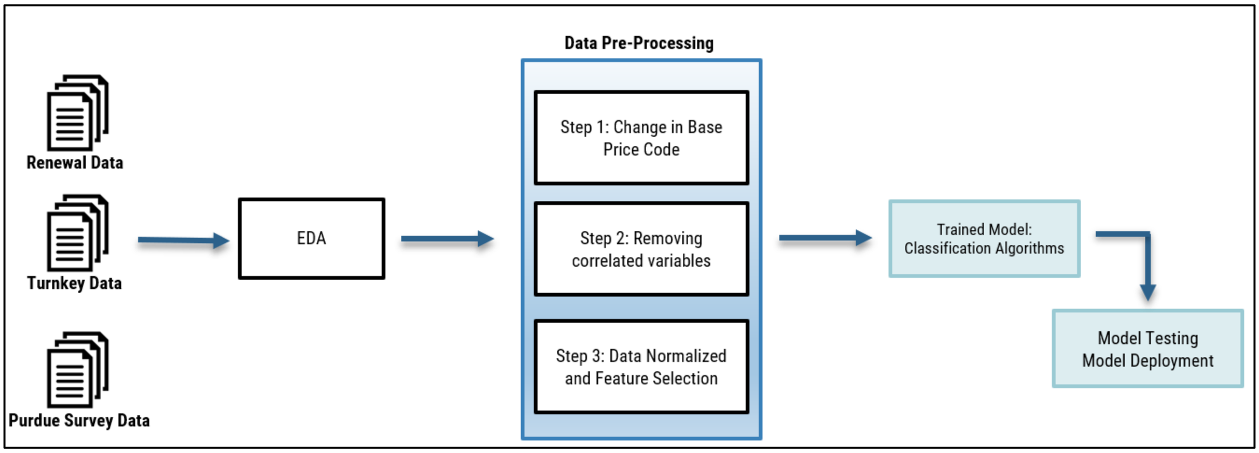

Renewal Data
Turnkey Data
Purdue Survey Data
EDA
Data Pre-Processing
Step 1: Change in Base Price Code
Step 2: Removing correlated variables
Step 3: Data Normalized and Feature Selection
Trained Model: Classification Algorithms
Model Testing Model Deployment

Figure 5

Based on the Figure 5 we are following the below mentioned methodology:

Data collection – NFL Client
It is performed using excels provided by the client which is further loaded in Python environment. It consists of 3 different tables:

- Renewal table consists of information related to Season Ticket Members (STM) like 'Account ID', Tenure, 'Zip Codes', Seats, Prices, 'Renewal Date' etc.
- Turkney Data consists of information related to STM demographics and financials with fields like Gender, Age, Education, Household Income etc.
- Survey data consists information related to client's feedback from the Seasonal ticket members regarding their satisfaction level, would they recommend this to someone else, does the club meet the expectations, if their problems are resolved etc.

Data Merging

- The data is merged for modeling purpose by combining the Renewal and Turnkey tables
- Based on 3947 records merged, we can conclude that these many Season Ticket Holders have information related to their demographics and financials
- Considering 60% of the information is lost during the merging with Turnkey data, hence for modeling purpose we are considering only the renewal dataset

Exploratory Data Analysis & Data Pre-processing

- For the renewal table, there are approximately 2.5% records missing for CountyName, State and 'Distance from Stadium'. Furthermore, there are approximately 40% records missing in Renewal related columns which provides information regarded 60% of the STM have renewed their membership so far. For our analysis any completely missing records will be removed

- The major part of the Exploratory Data Analysis was done using Tableau and are described from Figure 1 to Figure 4
- Normalization is using the in-built normalization command in python.
- The data is divided into train and test data each having 10271 rows and 22 columns. Each row represents the customers and each column is a feature which is either unique or common to each customer. The training set has one additional column Renewal Rate which is the label showing either 0 or 1.
- From this we are able to identify that there are 3 float columns, 1 datetime column, 9 integer columns and objects.
- The float column represents the continuous variables. We have shown the distribution of the float columns. We'll use an OrderedDict to map the renewal rate to colors because this keeps the keys and values in the same order as we specify. From Figure 5 plots we can say if there is any significant difference in the variable distribution depending on the renewal rate.

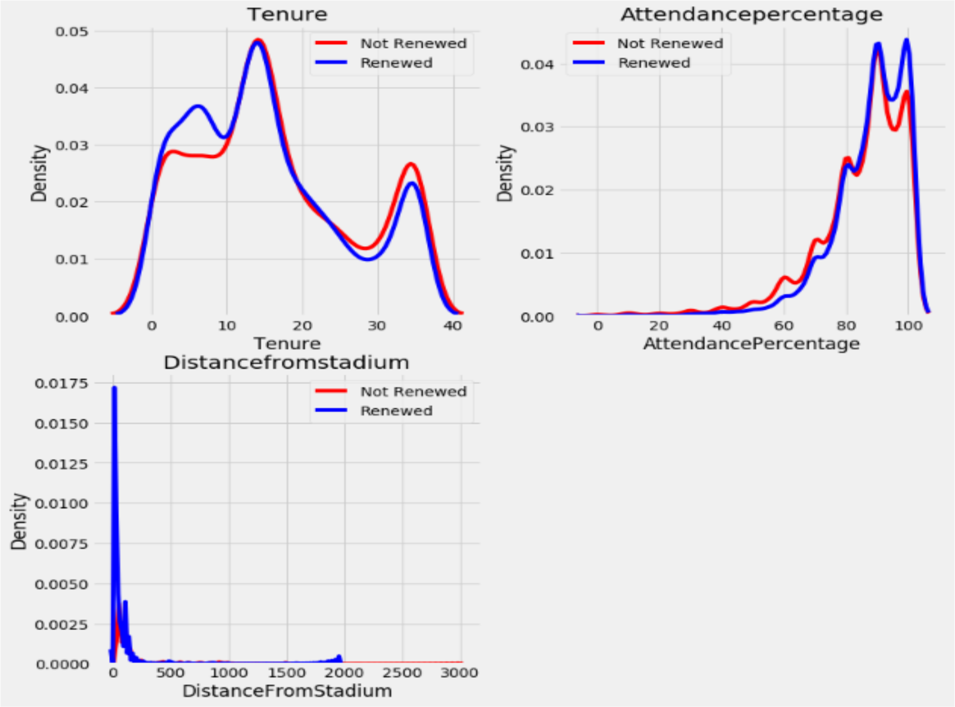


Figure 5

Exploring Label Distribution

- In order to understand the distribution of the label, Figure 6 is displayed to understand the balance using the training set.
- From the graph it is clear that we are dealing with an imbalanced classification problem. One problem with imbalanced classification problems is that the machine learning model can have a difficult time predicting the minority classes because it sees far less examples. One method to address this issue is to do oversampling.

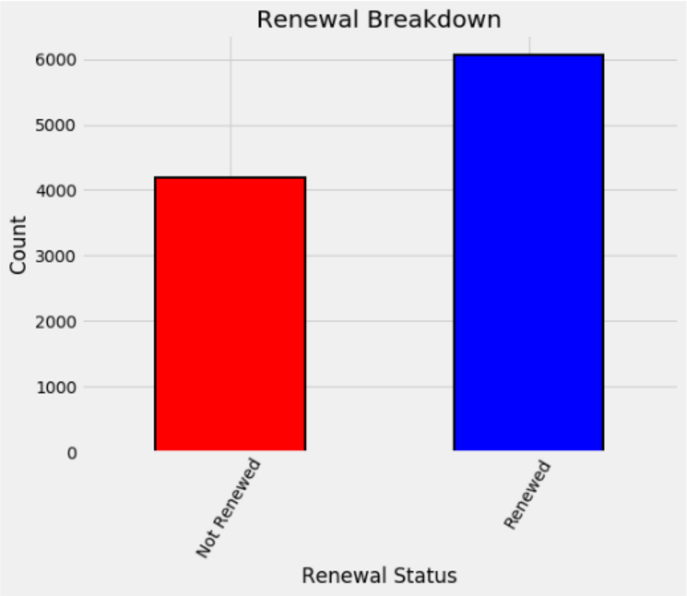


Figure 6

MODEL(s)

Metrics:

This is machine learning model that can predict renewal rate for the upcoming seasons. Our prediction is assessed by the standard F1 score. It is defined as the harmonic mean of precision and recall.

F_1=2*(Precision*Recall)/(precision+recall).

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

Assigning values:

From the data dictionary, we can understand that base price codes are classified from D to Z where D being the highest rate and Z being the lowest.

Feature Selection:

Feature selection is done using the FeatureTools package in python. From these 112 features are built. From these 112 features, 48 features are removed based on the following explained algorithm.

The algorithm chosen for feature selection is given as follows: Consider the dataset having K entities denoted as $E^{1\ldots.\ k.}$ The objective is to extract features for the target E^k. The entities with which E^k has both forward and backward relationship is denoted by E^F and E^B. The algorithm is described as follows:

```
1. function MAKE_FEATURES ($E^i$, $E^{1:M}$, $E_V$)
2. $E_V$=$E_V$ U $E^i$
3. $E_B$=BACKWARD ($E^i$,$E^{1\ldots M}$)
4. $E_F$=FORWARD($E^i$,$E^{1..M}$)
5. For $E^j$ ε $E_B$ do
           MAKE FEATURES ($E^j$,$E^{1\ldots M,}$$E_V$)
           $F^j$=$F^j$ U RFEAT($E^i$,$E^j$)
6. For $E^j$ ε $E_F$ do
           IF $E^j$ ε $E_V$ then
                   CONTINUE
           MAKE FEATURES ($E^j$,$E^{1\ldots M,}$$E_V$)
           $F^i$=$F^i$ U DFEAT($E^I$,$E^j$)
7. $F^i$=$F^i$ U EFEAT($E^i$)
```

This Feature selection algorithm is utilized by using featuretools package in python. (refer: *Deep Feature Synthesis: Towards Automating Data Science Endeavor (James Max Kanter, Kalyan Veeramachaneni)*

Models considered for prediction:

1. Random Forest:

   Random forest is a learning method for classification and regression done by classifying the features of importance by generating multitude of decision trees using the training set and gives output as the mean prediction or classification set.

2. Linear Support vector machine:

   Support vector machine is a supervised learning model for classification and regression. SVM builds a non-probabilistic binary classifier. There is gap obtained between the data points and the decision boundary of the classifier.

3. MLP classifier:

   Multi-Layer Perceptron model is a supervised learning model which has 3 layers of node: Input, Hidden and Outer layer. The difference between linear perceptron and MLP is presence of multiple layer and the non-linear activation function.

4. Logistic Regression:

   Logistic regression is statistical model which uses binomial regression for estimating the parameters present in the model. This has two levels 0 or 1 for response.

5. Ridge Classifier:

   It is formed based on ridge regression where there is a serious effect of multicollinearity present in the model. Multicollinearity affects the response as the value is far from the true value. Ridge induces a bias which helps in determining the correct value.

6. Linear Discriminant Analysis:

   Linear Discriminant analysis is used in machine learning and pattern recognition to find the linear combination of features that has two or more classes of response present in them. The dimensionality reduction is a major advantage of using this model.

7. K neighbors' classifier:

   K-neighbors is a non-parametric model for classification and regression. In the response is binary which is assigned by the class which has the common k nearest neighbors (k>0). It is otherwise known as lazy learning algorithm.

10-fold cross validation is performed on the models chosen and the F1 scores are compared to identify the best model.

RESULTS

The F1 scores are obtained for the method of cross validation for the chosen models are tabulated below:

Model	Precision	Recall	F_1 Score
Random Forest	0.71	0.684	0.717397
Linear Support Vector Machine	0.741	0.839	0.786287
MLP Classifier	0.75	0.77	0.754591
Logistic Regression	0.743	0.837	0.786109
Ridge Classifier	0.739	0.843	0.786959
Linear Discriminant Analysis	0.74	0.839	0.786219
K neighbors Classifier	0.698	0.757	0.724307

From this we could say that Linear Support Vector Machines as well as Logistic Regression has similar results in terms of F1 score except for SVM has comparatively better results. But **Logistic Regression can provide us with better interpretability** and hence is the best model out of these.

Since the data is normalized, the weights of the features determine their importance. A graph is shown displaying the 10 most important features present in the model.

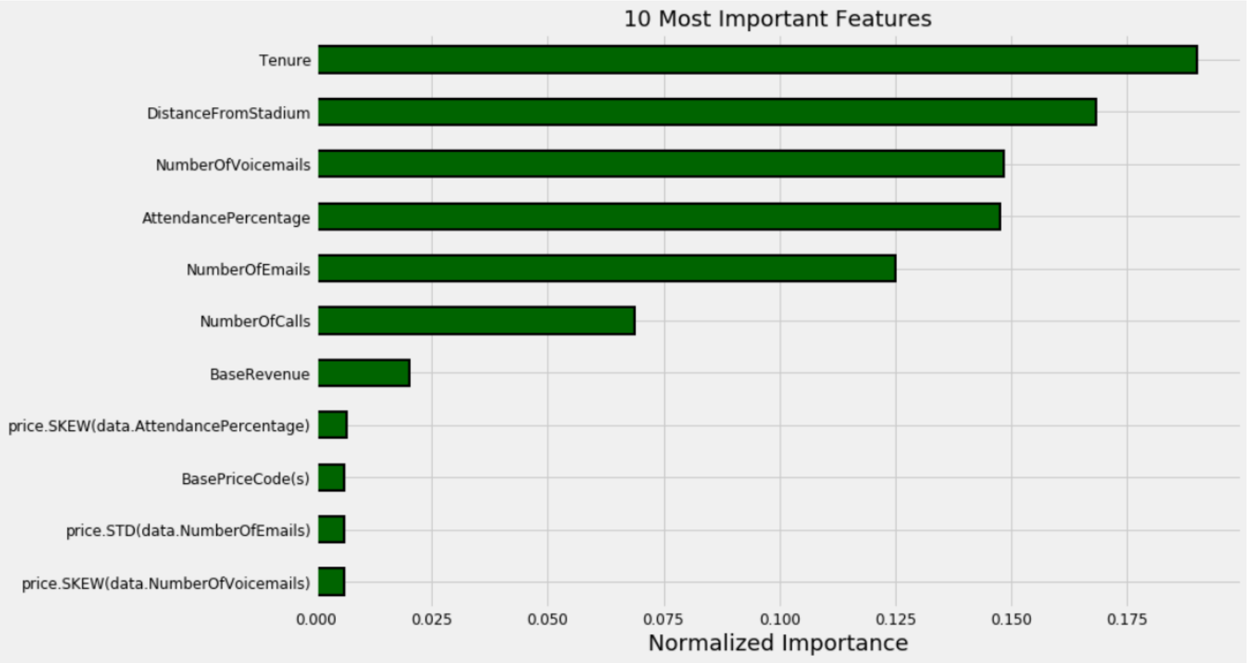
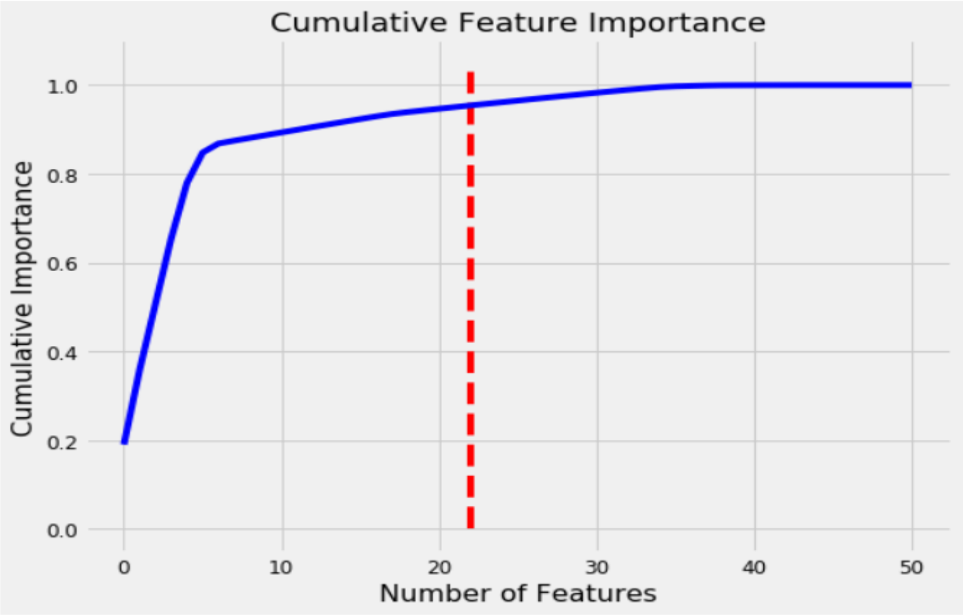


Figure 7

To understand the number of features to make accurate predictions, a cumulative plot is drawn which explains that 22 features are important for 95% cumulative importance in predictions.



```
22 features required for 95% of cumulative importance.
```

Figure 8

CONCLUSIONS

From this we could say that the Logistic Regression is the best prediction model. From the feature importance graph, we could say that the tenure, distance from the stadium, attendance percentage, communication with the customers and base revenue are most important features that determine the renewal rate. This is valid because each feature has serious implications to the ticket renewal and if any one of the features are affected there is huge impact present. Also, the model is trying to convey us the idea that if there's a constant interaction with the customers, then they would likely support the organization. If the Indianapolis Colts organization could bring necessary improvements to the prescribed features, they could see some good improvements in their season ticket renewal.

Improvement Areas: Considering only 40% of the turnkey data was available for the accounts present in renewal data which was a barrier. A lot of important variables related to Financial History and demographics can be utilized to better the accuracy of the existing model.