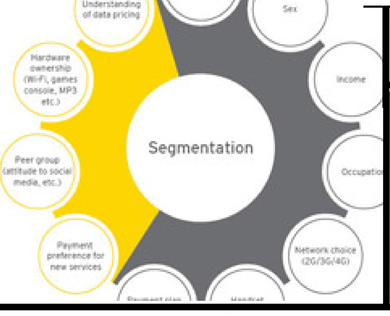


Telecom Segmentations:

Predicting Customer Gender Based on Network Behavioral Trends!

Augustus Madsen and Kyler Hart

Business Objective

How accurate will mobile telephony segments be in predicting a customer's gender?

Background

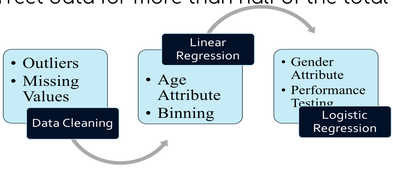
Mobile phone network marketers are responsible for segmenting customers based on behavioral patterns associated with mobile phone usage. These segments offer a diverse range of customer profiles, and provide information regarding professional and residential use, contractual data, and various network activity trends. By developing a regression model, marketers can predict with a certain accuracy the gender (Male or Female) of a given customer based on the behavioral data they have collected.

Data Description

There were 55 total attributes in the dataset. 53 of these attributes were of the real type, while 2 were polynomial: customer ID and gender. The data set uses "F" and "M" to differentiate between male and female. In the analysis, gender was chosen to be the predictive label due to its polynomial type, which is required for logistical regression.

Data Preparation

Before deciding on logistic regression, the first technique involved using linear regression, with the "age" variable as the predictive label. Unfortunately, because age was a value with multiple decimal points, the predictive model was unsuccessful more often than it was successful. The next attempt at a solution involved binning the data; however, those results also proved unsuccessful. This project used logistic regression techniques to estimate a model for the telecom segmentation data. Logistical regression is a model designed for predicting polynomial data. This type of regression was used with the "gender" variable as the predictive label (Figure 1). After cleaning the data of outliers and missing values, a RapidMiner design was created. First, the Excel file is retrieved and then the data set was split for training and modeling. For the performance testing of the model, the ratio resulting in the highest accuracy was a 50/50 split of the data. Next, the Logistic Regression operator was implemented into the design. More specifically, the IRLSM logistic regression solver was chosen as that option garnered the highest predictive model accuracy. An Apply Model operator was then added for gender prediction (Figure 2), followed by the Performance operator producing the matrix (Figure 3) showing the accuracy and precision of the model. Although the results were far from perfect, a model was produced that showed correct data for more than half of the total dataset (63.71%).



Data Preparation, Modeling, and Evaluation

01

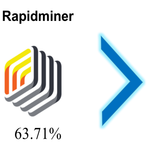
- Logistic regression proved capable of creating a model that could predict a customer's gender with an accuracy of 63.71%. However, this accuracy is not ideal as nearly 40% of all cases will be misrepresented. Two hypotheses were formed to explain the results. First, males outnumbered females in the dataset nearly 2:1. There were 2,051 males and 1,261 females in the sample. This large discrepancy could have skewed the model to produce a greater male percentage when compared to females. Secondly, the statistical results of the model displayed the confidence for both male and female predictions. Male confidence was 0.615 (61.5%) while female confidence was only 0.385 (38.5%). This shows that not only is the model predicting males more often, it is much more confident with the male choice it selected. This confidence disparity also explains the gender prediction gap.

02

- The segmentation application article showed that clustering methods carried the most significance in predicting an accurate model. The logistic regression used in this report was not as successful. Further analysis was attempted by using Watson Analytics, where the telecom dataset was imported and set to determine gender predictability. IBM Watson was only able to predict with a 62% accuracy (Figure 4), whereas the logistic regression predicted 63.71%. A decision tree was created using IBM Watson which displayed the key attributes used when predicting gender (Figure 5).

03

- While a model was able to be produced using Logistic Regression, there are more accurate ways of predicting within this data set. Clustering, as shown in the segmentation application article, is a much more effective model. However, this report found that logistic regression is one of many methods that can be used in this prediction. It is often necessary to evaluate all modeling options to determine the most efficient modeling technique for a given dataset.



Works Cited

- Chorianopoulos, Antonios. "Segmentation Application in Telecommunications." Effective CRM Using Predictive Analytics, Wiley, 2016.
- GmbH, RapidMiner. "Operators." Operator Manual - RapidMiner Documentation, docs.rapidminer.com/studio/operators/.
- Support." IBM Analytics Communities, 11 Apr. 2015, www.ibm.com/communities/analytics/watson-analytics/support/.

WKU GORDON FORD COLLEGE OF BUSINESS
BUSINESS DATA ANALYTICS