**Predictive Analysis on Astro First Movie Purchase**

|  |  |
| --- | --- |
| Problem statement:  Ability to create a campaign to target Astro customer to purchase Astro First movie based on historical behavior and profiles | |
| Business objective/goal:   1. To achieve high accuracy of prediction on targeted audience and purchase probability. 2. Reduce campaign costs | |
| Inferences (1):  Targeted customer by demographics, race | **Inferences (2):**  Probability of purchase |
| Inferences (3):  Purchase frequency/time band | **Inferences (4):**  Have particular Astro packages/services |
| Analysis Level: (Basic/Moderate/Extensive)  Moderate | **Accuracy of prediction method:**  Out-of-bag Accuracy of Random Forest on training dataset: **75.00%** |
| Algorithm/Method Used:  Random Forest | **Rationale to use this approach:**  As the dataset mainly consists of categorical variables, algorithms such as logistic regression, decision trees or its ensemble model - random forest, should be able to handle categorical values natively. I considered these algorithms, however I was concerned with the tendency of logistic regression and decision trees algorithms to overfit. Random Forest is an ensemble model, which combines a number of decision trees models. It usually yields better results than single models as it can detect different patterns in data and combining models would decrease the tendency to overfit the data. |
| Data Source Path:  <https://www.dropbox.com/s/s9z5016js4bsj1q/Predictive%20analysis.zip?dl=0> | |
| Data Sources -1:  Training data for Astro First Munafik.csv | **Source Description:**  Contains the customer demographics and purchase trends for training data |
| Data Sources -2:  Data to predict for Astro First Munafik.csv | **Source Description:**  Contains the customer demographics and purchase trends to create the campaign |
| Data Sampling:  152518 records | **Total projected rate:**  851830 record |
| Assumptions Made:   1. Munafik was typically purchased by Malays. 2. Male customers tend to purchase Munafik more than female customers. 3. Munafik was purchased mainly in the fourth quarter of 2015 and first quarter of 2016 (both Astro First and Astro Best) 4. Munafik was purchased by customers who do not have Value Pack or Super Pack, but have AOTG. |  |
| Result Summary:  Total number of users will buy: 580117  (The prediction result can be found under column “Purchased Munafik” in the attached file - Data to predict for Astro First Munafik - Submission.csv)  Total number of users probability to buy: 475969  (this is based on conditional probability according to customer demographics only [Race & Gender]). | **Detail Analysis Report Path:**  Git repository for the assignment:  https://github.com/shrrn/Astro-Assignment  Data Analysis:  <https://github.com/shrrn/Astro-Assignment/blob/master/astro_assignment_analysis.ipynb>  Test Data Prediction:  <https://github.com/shrrn/Astro-Assignment/blob/master/astro_assignment_prediction.ipynb> |
| Conclusion:  My approach for the predictive analysis task includes data visualization using plotting tools in Python in order to make sense of the data and to identify significant variables. As the response variable is categorical, Random Forest algorithm was chosen based on its suitability in analyzing categorical data and its performance as one of the top prediction algorithms. Random Forest reported 75% accuracy on the training data based on Race, Gender and Astro First and Astro Best customer purchases in the fourth quarter 2015 and the first quarter 2016. These variables are then used to make predictions to the test data, where 580117 customers are predicted to purchase the movie Munafik. | Name: Shireen Mohd Zaki  Date: 1st May 2017 |