Starbucks Capstone Challenge Proposal

1. **Domain background –** *The Starbucks project is coming out from customer marketing domain. Traditional marketing analytics or scoreboards are essential for evaluating the success or failure of organization's past marketing activities. But today's marketers or organizations can leverage advanced marketing techniques like predictive modeling for customer behavior, predictive lead scoring, and all sorts of strategies based on predictive analytics insights. Predictive techniques can make an organization marketing investment much more efficient and helps in regularly validating results. Connecting customer information to the operational data provides valuable insight into customer behavior and the health of your overall business.*

2. **Problem statement–***In this project, we would go through the predictive modeling technique for customer behavior through Starbucks dataset example. Starbucks provided simulated data that mimics customer behavior on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free). Some users might not receive any offer during certain weeks. Not all users receive the same offer, and this data set is a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks actually sells dozens of products.*

   In this project, Starbucks wants to connect offer data, customer data and transaction data (operational data) to gain insights about customer behavior and overall effectiveness of offers as a value for business.

   With the above key statement in mind, below is the main objective or problem motivation or problem statement:

Build a model that predicts whether a customer would respond to an offer or not.

Above problem is a classification problem and would be solved by using supervised machine learning algorithms.

3. **Datasets and inputs –** *The data is contained in three files:*
   - *portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)*
   - *profile.json - demographic data for each customer*
   - *transcript.json - records for transactions, offers received, offers viewed, and offers completed*
   - *Here is the schema and explanation of each variable in the files:*

portfolio.json

** id (string) - offer id*
** offer_type (string) - type of offer i.e. BOGO, discount, informational*
** difficulty (int) - minimum required spend to complete an offer*
** reward (int) - reward given for completing an offer*
** duration (int) - time for offer to be open, in days*
** channels (list of strings)*

profile.json

** age (int) - age of the customer*
** became_member_on (int) - date when customer created an app account*
** gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)*
** id (str) - customer id*
** income (float) - customer's income*

transcript.json

** event (str) - record description (i.e. transaction, offer received, offer viewed, etc.)*

** person (str) - customer id*
** time (int) - time in hours since start of test. The data begins at time t=0*
** value - (dict of strings) - either an offer id or transaction amount depending on the record*

4. **Solution statement –** *Supervised machine learning classifiers algorithms would classify if the customer would respond to an offer or not based on the offer data, customer demographic data and transactional data. We would prepare the combined and appropriate data with data analysis, visualizations and feature engineering to be fed into the best machine learning classifier to classify if the customer would respond or not.*

5. **Benchmark model –** *Our baseline model would predict that all users would respond to the offer. So, we will calculate f1_score of the baseline model against which we would compare our model to determine if our model is performing better than baseline model or not.*

6. **Evaluation metrics -** *For this case, evaluating a model with precision and recall would provide better insight to its performance rather than accuracy. Because, Starbucks would like to send offers to those customers whom have more chances of redeeming the offers rather than to send offers to all customers. F1-score metric is "the harmonic mean of the precision and recall metrics" and is better way of providing greater predictive power on the problem and how good the predictive model is making predictions. Refer* ***Classification Accuracy is Not Enough: More Performance Measures You Can Use*** for more information.

7. **Project design -** *The entire analysis would contain below steps:*

   1. Analyze each of the portfolio, profile and transaction data.
   2. Clean and transform each of the portfolio, profile and transaction data.
   3. Combine portfolio, profile and transaction data.
   4. Select a performance metric to analyze performance of the model and to compare different models.

5. Compute the performance of a baseline model against which performance of other different models would be compared.
6. Select best performing model based on the metric and training time.
7. Calculate the feature importances given by best estimator of the trained model.
8. Compute the performance of best model on test set and visualize the performance via confusion matrix plot.