Problem Statement

Introduction

Welcome! American Express Campus Analyze This is a first-of-its-kind Data Analytics competition by American Express®. Through this game, you will get a firsthand experience of the various facets of the exciting field of Data Sciences.

By the end of this week long, nerve-wracking, nail-biting, roller coaster ride, we are sure you would agree that Data Analytics is as addictive as gaming.

Gear up and Game On!

The sections below have details on the following:

Background

The Bank of Hoenn has decided to bet on data science this year. As a pilot project, they are looking for a robust analytical solution for their card offer extensions this month.

They have three kinds of card offers for their existing card holders:

1. **Supplementary Card:** This is a card with an additional balance over and above what one already has on his/her existing card. It is a privilege offered to the spouse, parents or children of the primary card holder. For this card, one needs to pay his/her entire monthly due in full, at the end of each month.
2. **Credit Card:** This is a typical credit card that allows the customer to pay his/her dues over a period of time, subject to interest being charged.
3. **Elite Card:** This is a premium card with high-end lifestyle benefits offered to customers who can afford its high annual fees. For this card, too, one needs to pay his/her entire monthly due in full, at the end of each month

With a fixed budget of $ 6,000 for this project, they need the best way to know which customers to call and for which offer. Each call costs them $ 6. A call to the right customer can be successful in the first try if the offer extended matches what the customer wants. If the customer is looking for a different card, he/she will inquire about it and end up taking up the offer. In that case the call costs $ 12 and is successful. If a customer who is called is not interested in any offer, the call costs $ 6 and is unsuccessful.

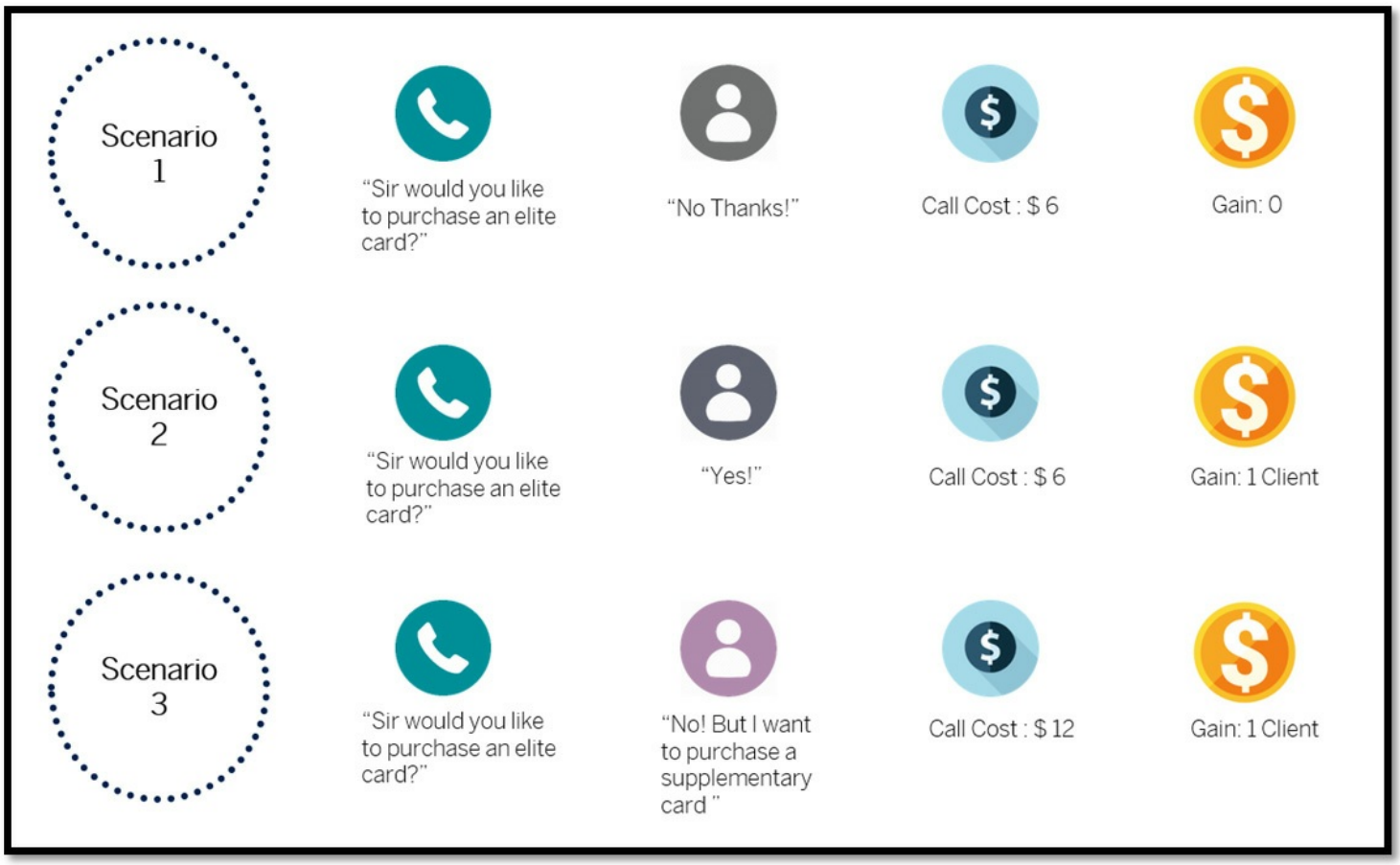
The bank has opened access to their past data and allowed vendors to compete and provide them the best solution.

Your analytics startup called **- Analyze This** has the responsibility to solve this through advanced analytics.

Information on **historical offer data** is what you have to create a list of customers to call , linked to the offer they should be called for.

Can your startup crack this puzzle? Do you have it in you to take Analyze This to the next level?

1. You have to create an ordered list of customers along with the specific card offers.
2. The order of the submission is the sequence in which the calls will go out.
3. Evaluation will stop when the budget is used up or when the list is exhausted.
4. There is no motivation to save anything from the $ 6,000 budget.



Data for Analysis

Following files can be downloaded for your analysis

1. **Training Dataset.csv**: This data has a sample of past card offer extensions. This data contains:
    a. Customer level data

b. Indicators for offers extended and indicators for success or failures of such offers

c. Demographic and spend variables

2. **Leaderboard Dataset.csv**: This data has historical customer data along with all the variables in the training dataset. The actual offers extended and the outcomes of the offer extensions are not present in this data.
3. **Final Dataset.csv**: This data has historical customer data along with all the variables in the training dataset. The actual offers extended and the outcomes of the offer extensions are not present in this data.
4. **Data_Dictionary.xlsx**: This sheet will give you the description of all the variables contained in the 3 datasets above.

Please note that you can make multiple submissions corresponding to the Leaderboard Dataset. However, for the Final dataset you can submit only one solution. For further details, please refer to the submission guidelines document available at the link below:
https://in.axpcampus.com/AnalyzeThis/submission_guidelines.php

Milestones

There are two milestones in this competition with exciting prizes. Check your mails regularly for more information about the milestones and keep an eye on the web site too!

Tips on Data Analysis

Following are some tips for the amateurs on how you can approach this data analysis game.

Any exercise in the field of data analytics would start with understanding the data. So, start off by understanding the datasets and descriptions provided to you.

Once you are familiar with the data, try to answer these questions:

So this is Khalid Works

1. What all data do I have?
2. What all data is useful and what is junk?
3. How can I organize this data to solve my problem?

Then, try to build the variables on the training dataset, define dependent and independent variables and then start modeling on the Training Dataset. This could involve building multiple models for different dependent variables. Or different models for different aims. Be clear about the aim of the exercise. In this game, you have to identify which customer should be called for what offer.

Once you are satisfied with your model, use it on the Leaderboard dataset and come up with your estimates of offer extensions for each customer. Follow the submission guidelines and

upload your ordered set of customers with offer extensions. Your submission will be evaluated in real time and you can compare how well you have estimated against other participants.

Keep fine tuning your estimates by trying to increase your leader board scores. Once satisfied, use the same logic to create the submission data for the final dataset.

You can use any tool, write your own algorithms, and implement any predictive modeling/Data analysis methods you may want to. For your final submission, you will have to provide details of the techniques you have used.

Popular Data Analysis Techniques

1. **Regression:**
   Regression is a mathematical process used to find a function that closely fits a series of data. The analysis involves defining the function that minimizes the difference between the data point and the value predicted by the function. There are several different techniques, the most common being by the method of least squares.

   For example, say you wanted to find an equation that dictated a certain stock's performance. You could take the closing price of that stock for every day in the last year. You then would be trying to figure out what equation satisfies all those points. The equation could be used to try to predict future performance.

2. **Logistic Regression:**
   Say, you want to figure out whether the stock price for a certain day would go up or not. You would again have the closing price of that stock for every day in the last year. We can do this using Logistic Regression. It gives you the probability of stock price rising.
3. **Support Vector Machine:**
   Imagine the previous scenario. In addition to closing price we have say some more indicators like volume traded as well, and we have a reason to believe that the price (as is often the case) is a complex function of these indicators. Then, to predict the upward or downward trends, SVM could be a better technique for the solution.
4. **Neural Networks:**
   Again, referring to the previous example, let's say, that we have certain indicators which are themselves complex functions of several different variables, and suppose we want to use them for the final prediction. In such a scenario, neural networks may give a better solution.

   A point to note, as we go down this hierarchy we might end up over fitting the data.

5. **Clustering algorithms:**
   Clustering algorithms are used in search engines that try to group similar objects in one cluster and the dissimilar objects far from each other. It provides result for the searched

data according to the nearest similar object which are clustered around the data to be searched.
As an illustration, Google uses clustering algorithms to classify different contents as News by parsing though the matter and examining the keywords.

6. **Recommendation engines:**
   Amazon/Flipkart/Netflix use collaborative filtering for recommendation.
   In essence, the algorithm represents each customer as a vector of all items on sale. Each entry in the vector is positive if the customer bought or rated the item, negative if the customer disliked the item, or empty if the customer has not made his or her opinion known. Most of the entries are empty for most of the customers. The algorithm then creates its recommendations by calculating a similarity value between the current customer and everyone else.
7. **Naïve Bayesian Text Classifier**:
   The best known use of Naïve Bayesian classification is spam filtering. It is a probabilistic classifier based on Bayes' theorem.
   For example, Emails use Bayes' formula for calculating the probability of an email to be classified as a spam, given already existing spams. This can be done by calculating probabilities associated with each word of the text to be classified as a spam.