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 Science.

By the end of this 5 day 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:

1. Background
2. Problem Statement
3. Data for Analysis
4. Milestones
5. Tips on Data Analysis
6. Popular Data Analysis Techniques

Background

The students of Ivyleague University have decided to embark on a start-up journey. They decided to establish and operate a departmental store named Tabz Departmental Store. This store offers a wide variety of products and services that caters to the daily needs of all the people staying in the city in which the university is located. Apart from Tabz, there are a couple of other departmental stores in the city. In order to tackle the intense competition, Tabz decided to offer its products and services on credit i.e., customers can consume products and services without paying for it at the time of consumption. Payment for the products and services consumed from Tabz during the month can be made at the end of the month. This is a revolutionary idea when compared to the traditional mode of cash payment and is expected to simplify the lives of people in the city while giving competitive advantage to Tabz. So, the founders decided to launch a co-branded credit card in association with a local bank, Banco. Tabz has decided to offer two types of cards:

a. Charge card: The balance is required to be paid in full each month

b. Lending card: Lending cards allow the customer to pay the balance over a period of time subject to interest being charged

An individual can apply for only one of the two types of credit card on offer. In order to extend the credit card to the individuals, Banco must first underwrite the applicant. Underwriting is the process by which the lender decides whether an applicant is creditworthy and should receive a credit line. Given the innovative business model of Tabz and the sound reputation of its founders, thousands of residents in the city submitted their application forms for the co-branded credit card from Tabz. Along with the data present in application forms, Banco also has access to the consumer bureau. Bureau is an agency that aggregates consumer borrowing and payment information for the purpose of assessing credit-worthiness of an individual and setting a limit on the cumulative credit that can be extended to an individual by lenders.

Banco has hired you to help underwrite each applicant and predict the credit worthiness of an individual. Banco has provided you with the customer application and bureau data with the default tagging i.e., if a customer has missed cumulative of 3 payments across all open trades, his default indicator is 1 else 0. Data consists of independent variables at the time T0 and the actual performance of the individual (Default/ Non Default) after 12 months i.e., at time T12. Banco’s expectation from you is to predict if an applicant will go default in next 12 months from the time of application submission.

At the time of launch of credit card, Banco earmarked a budget of $ 50,000 for processing the applications. After 12 months, Banco is keen to know if it had processed the right applications. If you predicted that the applicant will not default and the applicant actually does not default after 12 months, then Banco assigns a retrospective processing charge of $ 5. In every other case, Banco assigns a charge of $ 10.

Problem Statement

You have to create a list of applications in the order in which Banco should process them. With an objective to maintain healthy financials, Banco would like to process least risky applications first. Against each application, you also have to provide your prediction of default - 1 or 0, where 1 indicates a default and 0 indicates no default.

Assume:

- A resident of the city can submit only a single application form
- None of the applications submitted are fraudulent
- State any other assumptions in your final submission

Data for Analysis

Following files can be downloaded for your analysis.

1. **Training_dataset.csv**: This dataset contains:
    a Applicant level historic credit history
    b Performance in terms of default tagging i.e. 1 for default and 0 for no default
    c Application and bureau data
2. **Leaderboard_dataset.csv**: This data has historical applicant level data along with all the variables in the training dataset. The actual performance i.e. default tagging is not present in this data.
3. **Evaluation_dataset.csv**: This data has applicant level data along with all the variables in the training dataset. The actual performance i.e. default tagging is 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:
http://in.axpcampus.com/AnalyzeThis/submission_guidelines.php

Tips on Data Analysis

Following are some tips for the uninitiated 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:

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.

Once you are satisfied with your model, use it on the Leaderboard dataset and come up with your estimates of default for each applicant. Follow the submission guidelines and upload your predictions. Your submission will be evaluated in real time and you can compare how well you have estimated against other participants.

Keep fine tuning your predictions to increase your leader board scores. Once satisfied, use the same logic to predict if the applicant will default in 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.