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 3 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](https://in.axpcampus.com/AnalyzeThis/campusactivity/problem-statement.php#Background)
2. [Problem Statement](https://in.axpcampus.com/AnalyzeThis/campusactivity/problem-statement.php#ProblemStatement)
3. [Data for Analysis](https://in.axpcampus.com/AnalyzeThis/campusactivity/problem-statement.php#DataforAnalysis)
4. [Milestones](https://in.axpcampus.com/AnalyzeThis/campusactivity/problem-statement.php)
5. [Tips on Data Analysis](https://in.axpcampus.com/AnalyzeThis/campusactivity/problem-statement.php#TipsonDataAnalysis)
6. [Popular Data Analysis Techniques](https://in.axpcampus.com/AnalyzeThis/campusactivity/problem-statement.php#PopularDataAnalysisTechniques)

Background

XYZEE, a multinational financial services company decided to launch a premium credit card for businesses. Businesses use credit card as a payment enabler as well as to meet short term credit needs. XYZEE decided to offer this new card to its existing business card holders only. The credit card will offer a ton of benefits such as:

* Access to airport lounges
* 5X membership reward points
* Elite status at premium hotels
* Access to premium golf courses
* Extensive coverage across geographies

Given the benefits of the credit card, thousands of existing customers submitted their application forms for the credit card. Along with the data present in application forms and customers internal performance on their existing credit card relationships with the XYZEE, XYZEE also has access to the consumer bureau. Bureau is an agency that aggregates consumer borrowing and payment information which can be leveraged in assessing credit worthiness of an individual and setting a limit on the cumulative credit that can be extended to an individual by lenders. In order to extend the credit card to the individuals, XYZEE has to perform Initial Line Assignment.

Initial Line Assignment refers to the process through which credit line is­­ assigned to a new approved card lending account. The objective of initial line decision is to profitably accommodate cardholder’s spending and revolving needs while controlling the amount of write-off, thus ensuring the financial health of the card portfolio and card-member satisfaction.

Line, also referred to as a Credit Limit, is one of the most important features of a credit card. Line is communicated to the customer and allows them to manage how much they can borrow on a single credit card account.

Line is a key component of the economics of a lending business. It drives profitability and customer satisfaction.  
Line impacts profitability in the following ways:

* Higher lines can drive higher spend and balance
* Higher lines will drive higher write-offs

XYZEE has hired you to predict with level of line category (low, medium, High) for each applicant as per its need and credit worthiness. XYZEE will leverage this categorization in order to come up with the actual lines for each application. XYZEE has provided you with the customer application, customer internal performance data with line categorization on existing cards. XYZEE’s expectation from you is to predict the level of credit line of an applicant in terms of low, medium and high.

Customer internal performance data comprise of spend on existing cards, line on existing card, predicted risk for existing card accounts, revolve balance etc.

Problem Statement

You have to provide the line level you wish to assign against each card application. You have to assign line level to the applicants from the below given 3 line level options only:

1. Low
2. Medium
3. High

Assume:

* A customer can submit only a single application form
* None of the applications submitted are fraudulent
* State any other assumptions in your final submission
* Low line level indicates low need/ creditworthiness
* Medium line level indicates medium need/ creditworthiness
* High line level indicates high need/ creditworthiness

Data for Analysis

Following files can be downloaded for your analysis.

1. **Training\_dataset.csv:**: This dataset contains:
   1. Applicant level historic credit history
   2. Existing credit line on credit card
   3. 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 credit line 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 credit line 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: [Guidelines and Submission](https://in.axpcampus.com/AnalyzeThis/campusactivity/guidelines-and-submission.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.

1. **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.

1. **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.

1. **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.

1. **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.

1. **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.

1. **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.