Welcome **Telugu Knights**

Logout

Problem Statement

Download Files

Data_Dictionary.xlsx | Training_Dataset.csv | Leaderboard_Dataset.csv | Final_Dataset.csv | 2015_Submission_Deck.pptx

Welcome! American Express Campus Analyze This is the third edition of a first-of-its-kind Pan-IIT 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 10 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

Background

The Island of Hoenn is gearing up for upcoming polls. Citizens are waiting with bated breath as news agencies reveal their predictions on which party is likely to emerge victorious.

Much to the disappointment of all the citizens, there is discrepancy in these poll predictions amongst news channels.

So many inconsistent predictions didn’t go down well with an inquisitive bunch of students. Wondering how difficult it might be to crack it, comes their 'Eureka' moment.

An idea to create their own ‘Start Up’ to analyze poll sentiments and predict the winner. A start up called - Analyze This.

They gather data for a sample of citizens of Island of Hoenn and get started.

Information on historical voting pattern, rally attendance, party agenda and demographics is what they have at hand to predict the winner amongst the 5 competing parties.

Can you help these students crack this puzzle? Do you have it in you to start your own Analyze This?

Problem Statement

Based on the data the students above have collected, you have to:

1. Identify which citizen is likely to remain loyal and vote for the same party as they voted for in the last polls.

2. Identify which citizen is likely to switch votes and NOT vote for the same party as they voted for in the last polls.

3. Identify overall winner of the polls.

Data for Analysis

The following files can be downloaded for your analysis

1. Training_Dataset.csv: This data is for historical votes and voting preference of a sample of citizens. The data has information on:

   a. Past history of voting for all citizens

   b. Who they will vote for in upcoming polls

   c. Donation, Rally Attendance, Demographics of these citizens

2. Leaderboard_Dataset.csv: It has information on the historical vote for all citizens along with donation, rally attendance and demographics.

3. Final_Dataset.csv: It has information on the historical vote for all citizens along with donation, rally

attendance and demographics.

4. Data_Dictionary.xlsx: This sheet will give you the descriptions of all the variables contained in the 3 datasets above.

Please note that the Leader board data submissions are restricted to only 10 submissions per day per team and 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/campusactivity/guidelines-and-submission.php

Clues and Milestones

During the game 1 clue will be released in order to help you solve the problem better.

This clue has the potential to give an extra boost to your solutions provided you can use them to your advantage. Check your mails regularly for information about the clues and keep an eye on the web site too.

We have defined 1 Milestone on the scores calculated for the Leader Board Submissions.

The first 2 teams to cross a milestone will be awarded. The milestone will appear on the leader board on the site.

Milestone Score: 600,000

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. You need to match the citizen’s choice of vote.

Once you are satisfied with your model, use it on the Leaderboard and come up with your estimates of which poll results for each citizen. Follow the submission guidelines and upload your estimates. Your submission will be evaluated 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. Keep an eye on the clues to better your solution. Once satisfied, use the same logic to estimate the vote preference of citizens 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.