**Gazing into the Data Abyss - Loan Prediction**

*Hindsight: For the complete code and steps taken with the data analysis, kindly look at the code at* [*https://www.kaggle.com/shawnd29/gazing-into-the-data-abyss-loan-prediction*](https://www.kaggle.com/shawnd29/gazing-into-the-data-abyss-loan-prediction)

Stories. This is the biggest way that we have delivered ideas from one generation to the next. As we look at the diversity of information that we uncover, it seems that everything that has been accomplished so far stems from our need to take risks and brave it all.

With this in mind, loans are one of the quickest way that people leverage their assets to obtain a commodity that is temporarily beyond their grasp. However, there has to be a grounded reason why said amount is borrowed and the credibility of paying back this amount.

This is currently being estimated by the credit industry and the truthfulness of a candidate is carefully scrutinized. The main aim behind this project is to uncover these insights and bring to light what are the most important features that credit-based enterprises keep in mind when an individual requests for a loan.

NOTE: This dataset was acquired from the Loan Prediction Practice Problem (Using Python) by Analytics Vidhya and a detailed step by step guide can be found at:

<https://courses.analyticsvidhya.com/courses/loan-prediction-practice-problem-using-python>

So let is jump into this story and uncover the insights:

**First Steps**

Our initial steps take us to extract and read the data and get a basic outline of the model.

**Categorical Independent Variable vs Target Variable**

Now, let is take a granular look at the data:

## Numerical Independent Variable vs Target Variable

A numerical approach our data can be seen below:

## Working on the Model

Now that the data is cleaned, we can focus on the model. Some the steps that e have accomplished are:

* We convert the categorical values through dummy values via one-hot encoding
* Split the train-test with a 70-30 ratio
* Performing a logistic regression

**Validation methods - Stratified K-folds**

In terms of validating data, stratified K-Folds are one of the various methods that are used. This is to ensure each class is (approximately) equally represented across each test fold (which are of course combined in a complementary way to form training folds)

Since there is a disparity between the 'Yes' and 'No' labels of the Loan status, this makes it mode distinct approach to follow.

**Using Domain Knowledge**

Based on the data, we derive the following values:

1. Total Income - The sum of the Applicant Income and Coapplicant Income. If the total income is high, chances of loan approval might also be high.
2. EMI - EMI is the monthly amount to be paid by the applicant to repay the loan. People who have high EMI’s might find it difficult to pay back the loan. We can calculate the EMI by taking the ratio of loan amount with respect to loan amount term.
3. Balance Income - This is the income left after the EMI has been paid. Idea behind creating this variable is that if this value is high, the chances are high that a person will repay the loan and hence increasing the chances of loan approval.

Ultimately, let us consider the feature importance

## Outcomes and Takeaways

The main reason for a consumer to take a loan is to acquire temporary monetary leverage in order to purchase commodities beyond their current spending range. Conversely, the federal system has the right to accept or deny said consumers from a loan request.

The key takeaway from this data shows a good credit history will go a long way to request a greater loan from the bank. This stands to reason as lenders will know that said consumer is reliable enough and will be more likely to pay back their accrued loans

With all that has been said, this project focused on the analysis that goes behind investment and credit-risk organizations. The insights covered here will go a long way to uncover more in the financial perspective.