

Credit Card Lead Prediction

Muskan Gulati | JOB-A-THON | 30th May, 2021

# Overview

For any Business, it’s important to understand customer requirements. Machine Learning allows companies to tailor products to consumer needs. How to accurately predict customers’ adoption behavior is becoming more important to many credit card marketers as competition increases. So how do credit card issuing banks navigate this scenario of consolidating rapid growth today while preparing for the future? The answer lies in adopting a more customer centric approach. If issuers are able to better predict consumer behavior, they have a stronger chance of offering highly personalized solutions. Adoption of artificial intelligence and more specifically machine learning offers this opportunity. How it can be done?

Let’s understand considering below Business case:

## BUSINESS CASE

Happy Customer Bank is a mid-sized private bank that deals in all kinds of banking products, like Savings accounts, Current accounts, investment products, credit products, among other offerings.

The bank also cross-sells products to its existing customers and to do so they use different kinds of communication like tele-calling, e-mails, recommendations on net banking, mobile banking, etc.

In this case, the Happy Customer Bank wants to cross sell its credit cards to its existing customers. The bank has identified a set of customers that are eligible for taking these credit cards.

Now, the bank is looking for your help in identifying customers that could show higher intent towards a recommended credit card, given:

* Customer details (gender, age, region etc.)
* Details of his/her relationship with the bank (Channel\_Code, Vintage, 'Avg\_Asset\_Value etc.)

## dATASET dESCRIPTION

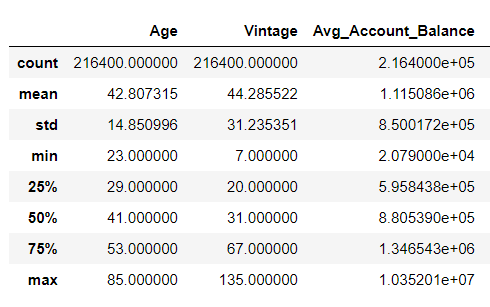
The dataset consists of 245725 observations and 10 features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Field Description | Variable Type | String/ Numeric | Ordinal/Nominal |
| ID | Unique Identifier for a row |  |  |  |
| Gender | Gender of the customer | Categorical | String | Nominal |
| Age | Age of the customer | Continuous | Numeric |  |
| Region\_Code | Code of the region for the customer | Alphanumeric |  |  |
| Occupation | Occupation type for the customer | Categorical | String | Nominal |
| Channel\_Code | Acquisition channel code for the customer | Categorical | String | Nominal |
| Vintage | Vintage for the customer(In Months) | Continuous | Numeric |  |
| Credit\_Product | If the Customer has any active credit product (Home loan, Personal loan, Credit Card etc.) | Categorical | String | Nominal |
| Avg\_Account\_Balanace | Average Account Balance for the Customer in last 12 Months | Continuous | Numeric |  |
| Is\_Active | If the Customer is Active in last 3 Months | Categorical | String | Nominal |
| Is\_Lead(Target) | If the Customer is interested for the Credit Card  0 : Customer is not interested  1 : Customer is interested | Categorical | Categorical | Nominal |

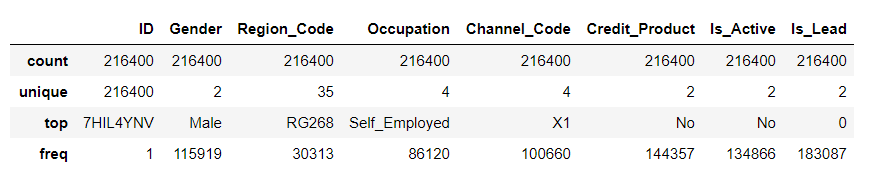
## Data Preprocessing and Exploratory Data Analysis

Before implementing machine learning algorithm it is important to understand the original data and various trends that data is following.

**Descriptive Statistics**: Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. With descriptive statistics you are simply describing what is or what the data shows.



The above table shows descriptive statistics for numerical data type. It returns mean, std, count and IQR values. The little difference between mean and median of the numerical variables show that there is little or no skewness in these variables.



The above table shows summary for object data type. It returns the count, unique value it contains, the value with higher frequency.

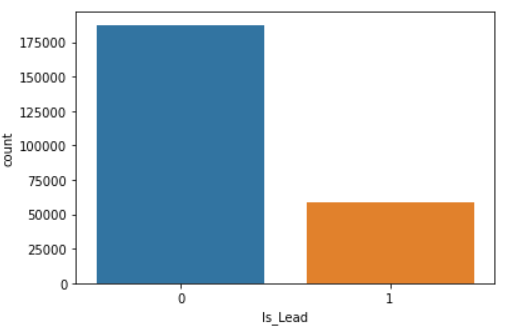
**Data Cleaning**:

**Detecting missing values:** Our dataset consist of 29325 NaN values in Credit\_Product. Since the number of missing values is large, we chose to replace those values with “Unknown” so that we don’t miss any important data.

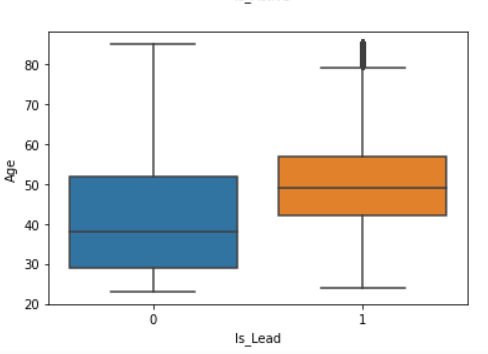
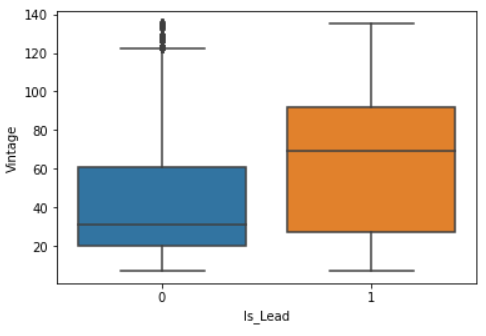
**Data Visualization**:

Inference from data visualization:

* The distribution of Is\_Lead is highly imbalanced i.e. out of customers are interested in credit cards which is approximately of total number of customers.

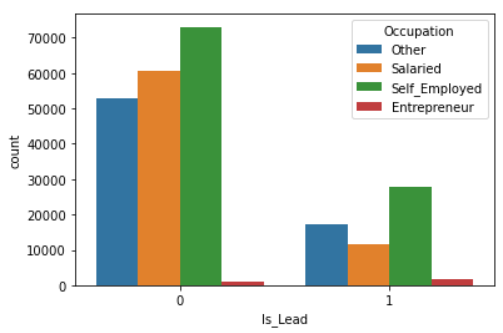
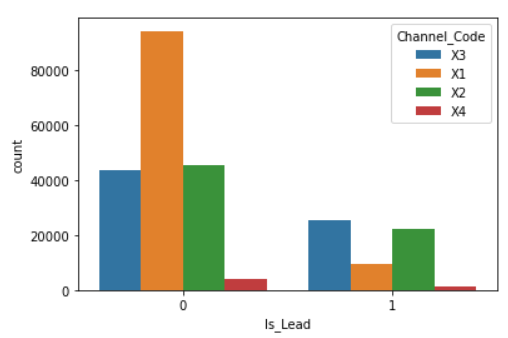


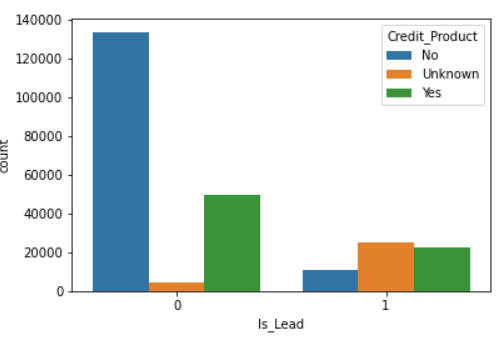
* Distribution of Is\_Lead against continuous variables

The above boxplots shows how target variable changes with different variables:

* The customers with more age are interests in credit cards.
* Customers having more vintage are interested in credit cards.
* Distribution of Is\_Lead against Categorical variables:



Inference made from above count plots:

* Salaried customers are less interested in credit cards.
* Customers with channel code X1 are less interested in credit cards.
* Customer who does not have credit product are least interested in credit cards.
* Customers for which we don’t know whether they have credit products or not are more interested in credit cards.

**Hypothesis Testing:**

A statistical hypothesis is a hypothesis that is testable on the basis of observed data modeled as the realized values taken by a collection of random variables.

These models includes statistical tests and make inferences about the population.

Below mentioned are the conclusions drawn from these statistical tests:

* The customers who are not interested in credit cards have average age less than the customers who are interested.
* The average vintage months of customers who are not interested in credit cards is less than the customer who are interested.
* Self-employed customers are most interested in Credit cards.
* Customers with channel code X3 are less interested in Credit Cards.
* Customers for which we have don’t know whether they have credit products or not are most interested in credit cards.
* Customers who does not have active products are not interested in Credit Cards.
* Salaried customers have least vintage and customers who are entrepreneurs have more vintage.
* Customers with no credit product have average age less than customers with credit product and unknown credit product
* Customers with unknown credit product have more vintage
* Customers who are active have average age more than the customers who are not active
* Customers who are active have average vintage more than the customers who are not active.

**Predicting Attrition** :

Since our main aim is to predict whether the customer is interested in Credit Cards or not. We have performed a classification based models: Random Forest.

Before digging deeper, we will understand the basic steps used before applying any classification model to the data.

* Converting categorical variables to dummy variables: Predicting models require all input and output variables to be numeric. The categorical variable is therefore converted to dummy variables. A dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.
* Splitting data to train and test: In a dataset, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. So, we use the training data to fit the model and testing data to test it. The models generated are to predict the results unknown which is named as the test set.
* Balancing target data: As we have already seen that the distribution of target variable is highly imbalanced. When you have imbalanced classes, your model might just learn to choose the majority class instead of reacting to the data. So we have used SMOTE technique to balance the Is\_Lead distribution.

Now, let’s dive deeper into classification model:

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

**Results:**

|  |  |
| --- | --- |
| Parameter | Result |
| Accuracy score | **0.8582** |
| Precision | **0.77** |
| ROC\_AUC | **0.872** |

* The accuracy score is 0.8582 which means out of total predictions, the model is able to predict 85% of the times correctly.
* The precision is 0.77 which means when it predicts the customer is interested, it is, it is correct 77% of the time.
* ROC\_AUC score is 0.0.87 which signifies how good the model is aggregating all the threshold values.

## SUMMARIzation

The main aim of this project was to find the similar characteristics followed by customers who are interests in credit cards. Using data exploration, we did observe some of the common patterns. To increase credit card sales, the bank should target customers with age above 40 as they show higher intent towards credit cards. However, bank should survey customers with lower age about their requirements and provide customized solution. The customers with more vintage are more interested in credit cards. Salaried customers need more focus, they should be provided with schemes beneficial to them. Self-employed customers and customers having channel code X3 should be targeted as they are more interested in credit cards. Customers having no credit products are not interested in credit cards which can be justified as they are not aware of the services provided. They should be made aware of the services provided by the bank. As a suggestion, the bank can introduce a forum for the customers where they can share their views about their issues, why they are not interested, whether they are satisfied with the service sand bank can work on customized schemes. Since customers who are not interested in credit cards have no credit products, bank can introduce them about their services.