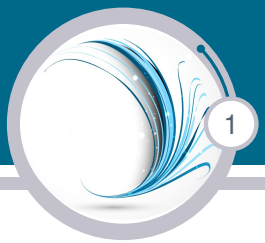
Advanced Data Science Capstone

MSc. Nicanor MAYUMU
nicanor@aims.edu.gh

May 17, 2020

Content



Data set - Value prediction
Data Source

Data set is from Kaggle Santander value prediction competition link.

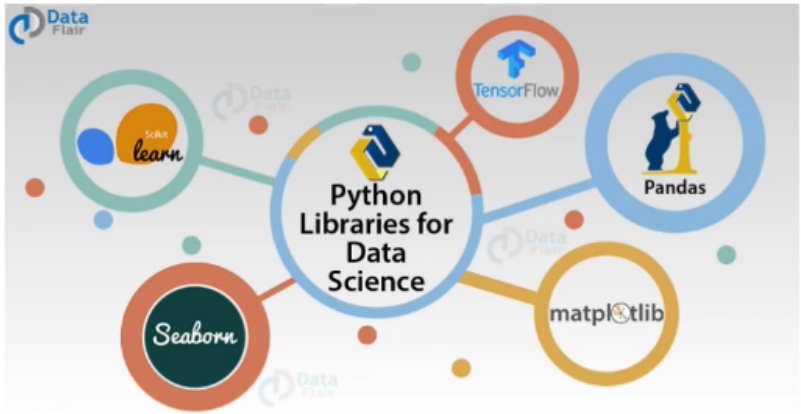
- August 13, 2018
- 4993 features
- The train set contains 4459 records
- The test set contains 49933 records

The data is anonymized. The task is to predict the value of target column in the test set using a machine learning algorithm.

Technology

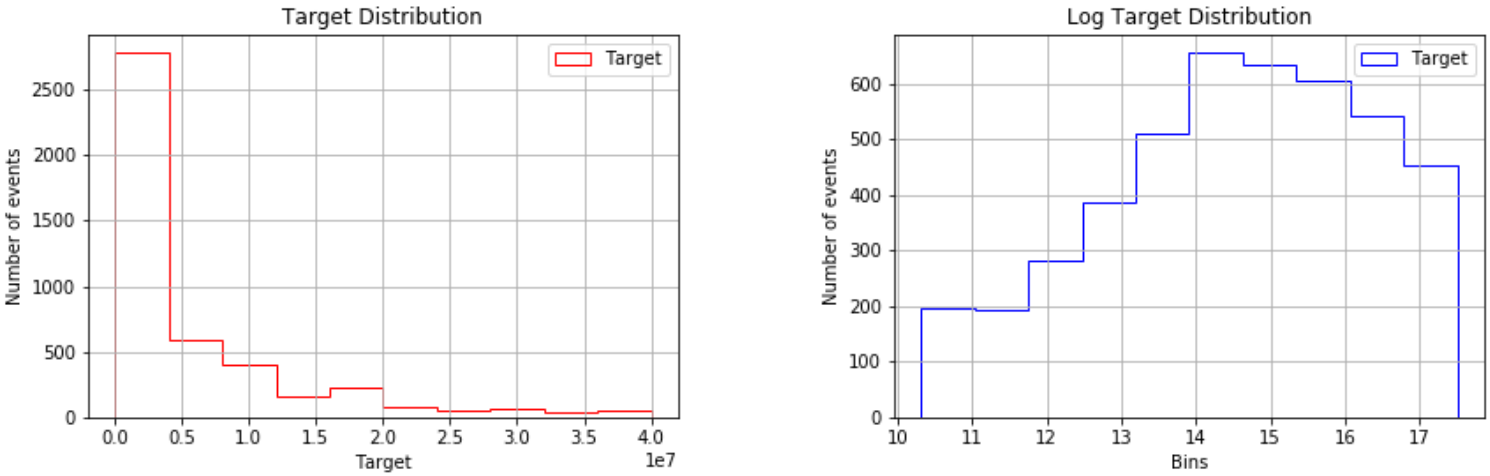
IBM Watson

- IBM Watson cloud with jupyter notebooks
- Python libraries for data science



Data assessment
Target variable

The target is skewed. We used log transform for transforming skewed data to approximately conform to normality.



(a) Target distribution

(b) Log target

Figure: Log Transformation of target

Data assessment
Data quality

Data analysis and visualization using pandas :

- Data types : float, double and object
- Overview of data set
- Checking for null entries, missing values...

Feature Engineering
Pearson Method, Backward Elimination

- The pearson correlation : None of features is highly correlated to the target variable.
- Backward Elimination : 40 selected features with p-values less than 0.05.

Performance metric
Root Mean Squared Error

Some of the standard evaluation metrics used to measure the performance of regression models:

- **Root Mean Squared Error**: RMSE is the square root of MSE. While Mean Squared Error is given by :

$$MSE = \sum_{i=1}^{n} \frac{(\omega^T x(i) - y(i))^2}{n} \tag{1}$$

It is the sum, over all the data points, of the square of the difference between the predicted and actual target variables, divided by the number of data points.

Performance metric
R-squared

Some of the standard evaluation metrics used to measure the performance of regression models:

- **The R-squared coefficient**: The R-squared coefficient, also known as the coefficient of determination, is a measure of how well a model fits a dataset.
- An R-squared coefficient generally takes a value between 0 and 1, where 1 equates to a perfect fit of the model.
- We split the data on training(90%) and validation(10%), then train on spark.

Selected Models

MODELS	RMSE	R^2	Data
Linear Regression	0.2904	0.9725	Original data
Gradient-boosted tree regressor	0.0698	0.9983	
Keras Regressor	0.40	n/a	
Linear Regression	0.3901	0.9727	Selected data
Gradient-boosted tree regressor	0.0698	0.9984	
Keras Regressor	0.30	n/a	

Table: Selected Algorithms

Summary

Finaly, the trained model can be used to predict target which is the customer value.

- collected data
- prepare and clean the data
- feed the data with a pre-trained model
- model predict the target on test set.