ANALYSIS OF HOME CREDIT GROUP’S CREDIT DEFAULT RISK

BY: NIRAV P. SHETH

INTRODUCTION

- CLIENT: HOME CREDIT GROUP
- PURPOSE: OPEN MORTGAGES SERVICES AVAILABLE TO THE UNBANKED POPULATION
- GOAL: CREATE PREDICTIVE MODEL TO DECIDE WHETHER TO PROVIDE THE LOAN
  - ATLEAST LESS THAN NATIONAL AVERAGE

DATA SETS

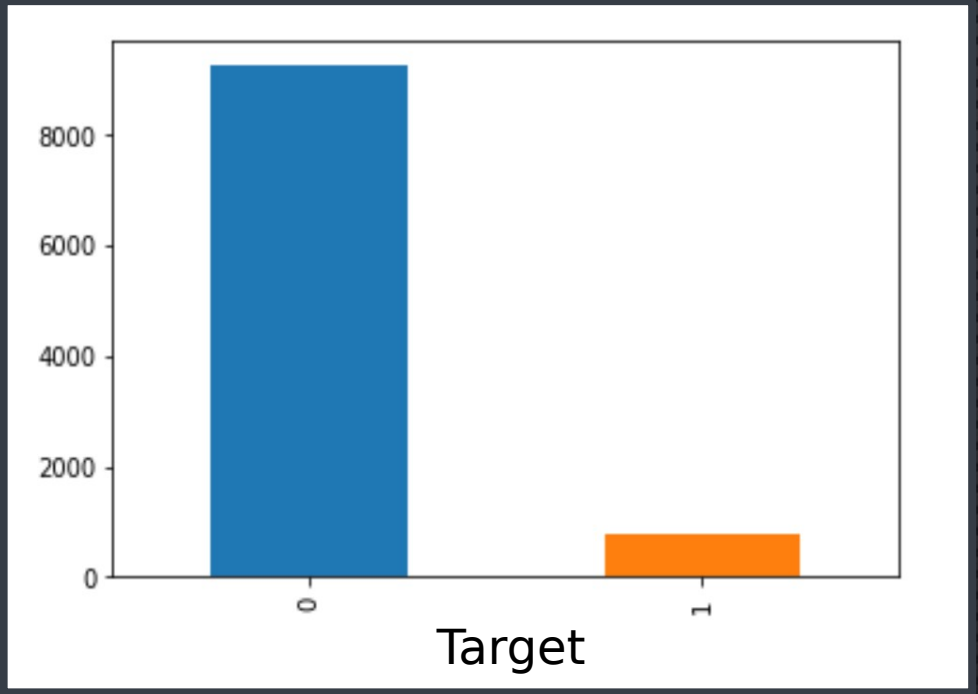
- TRAIN DATA SET:
  - 308K CLIENT APPLICATIONS WITH FINANCIAL, BEHAVIORAL, TIME-BASED, AND DESCRIPTIVE INFORMATION
- PREVIOUS CREDIT DATA:
  - PREVIOUS APPLICATION: INFORMATION IF CLIENT HAD PREVIOUS APPLICATION
  - 13.8M ROWS OF DATA FROM CREDIT BUREAU
    - 50% OF CLIENT APPLICATION AVAILABLE VIA CREDIT BUREAU
  - INSTALLMENT PAYMENTS: CREDIT HISTORY INFORMATION ON PAYMENTS ON LOANS
    - HELPS TO CHECK ANY NONPAYMENTS FOR CURRENT CLIENTS

DATA CLEANING

1. AGGREGATING
    1. HISTORICAL DATASETS NEEDED TO BE AGGREGATED DUE TO 1 TO MANY RELATIONSHIPS
    2. USED PIVOT_TABLE METHOD WITH NUMPY.MEAN, SUM, AND COUNT FUNCTIONS
2. MERGING
    1. LEFT MERGED APPLICATION TRAIN WITH AGGREGATED DATASETS
3. DUMMYING CATEGORICAL COLUMNS
    1. USED PANDAS GET_DUMMIES TO CONVERT CATEGORICAL COLUMNS INTO NUMERICAL
4. FEATURE SELECTION
    1. USED SELECTFROMMODEL WITH LASSOCV AS CLASSIFIER TO FIND TOP 15 COLUMNS

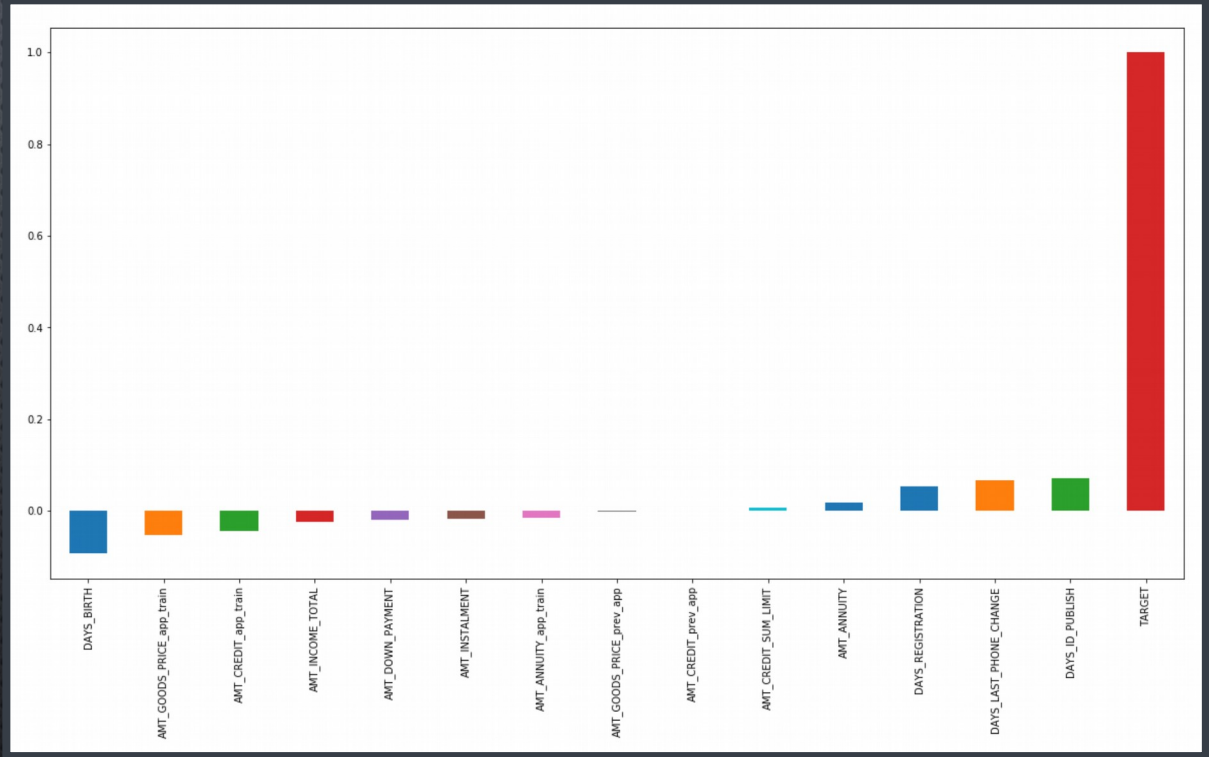
CLIENT APPLICATION: TARGET COLUMN

- TARGET COLUMN:
  - 0: PAYMENT ON TIME (93% OF THE SAMPLE)
  - 1: DELINQUENCY >30+ DAYS (7% OF THE SAMPLE)
- US AVERAGE DELINQUENCY RATE: 4.4%
- WHILE IT LOOKS LOW, IT IS STILL QUITE LARGE COMPARED TO NATIONAL AVERAGE

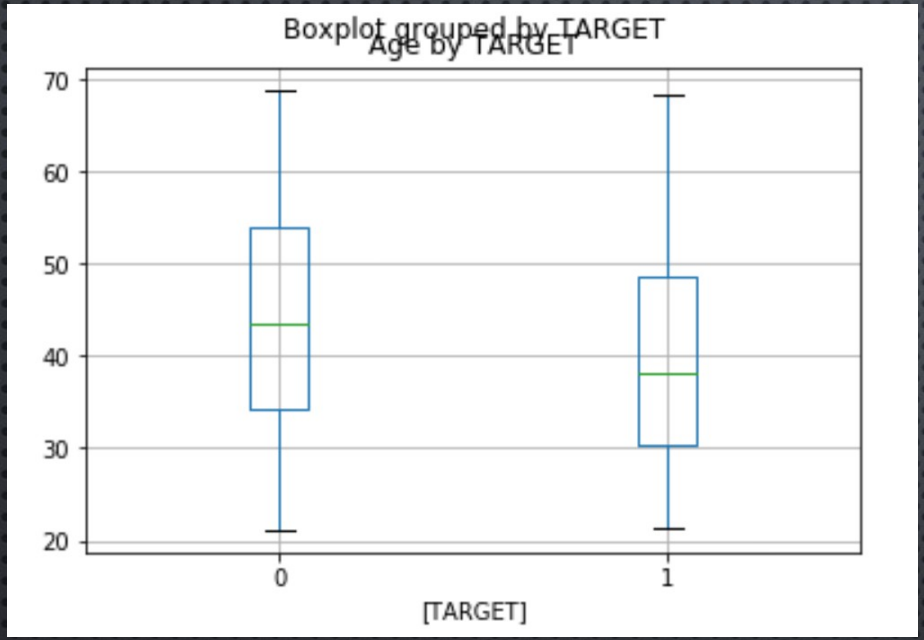


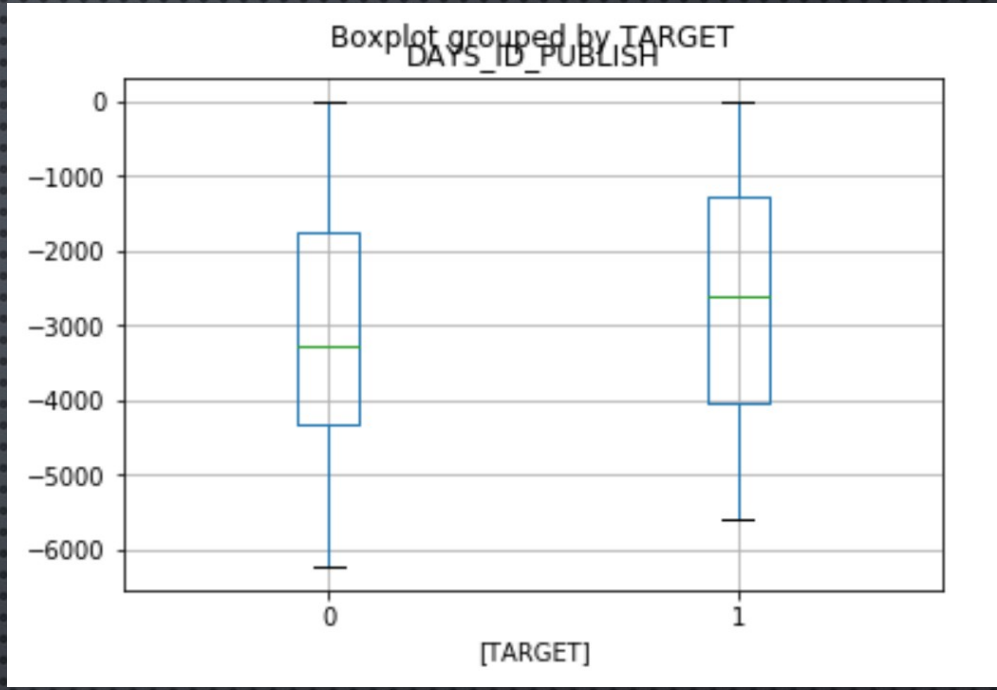
CORRELATION ANALYSIS VS TARGET

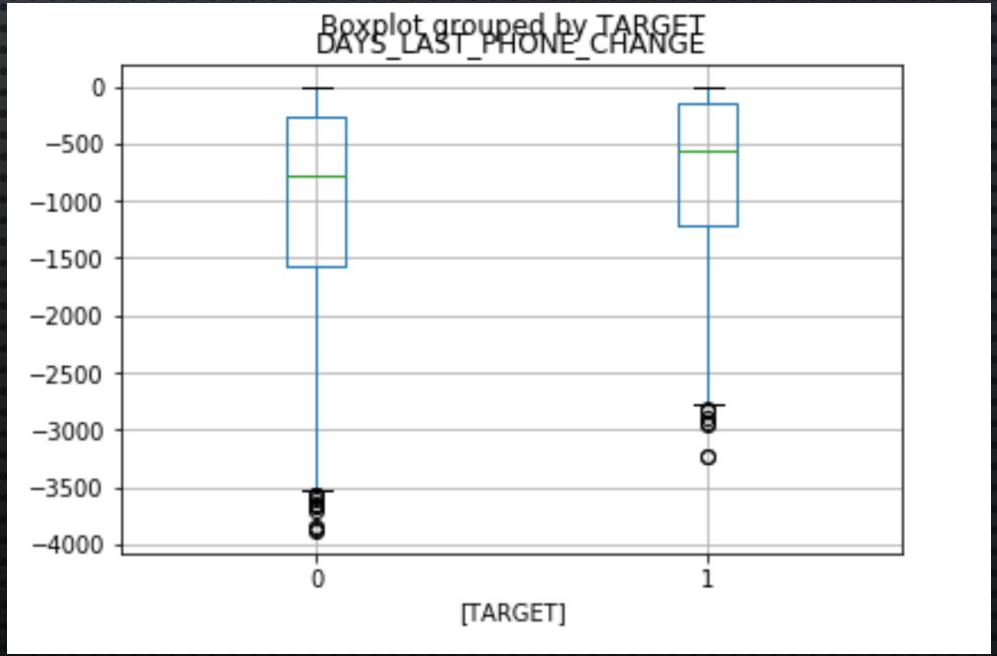
- 3 COLUMNS OF INTEREST:
  - DAYS_BIRTH: YEARS OLD
  - DAYS_LAST_PHONE_CHANGE: TIME SINCE LAST PHONE CHANGE
  - DAYS_ID_PUBLISH: TIME SINCE ID CHANGE



TARGET VS FEATURES







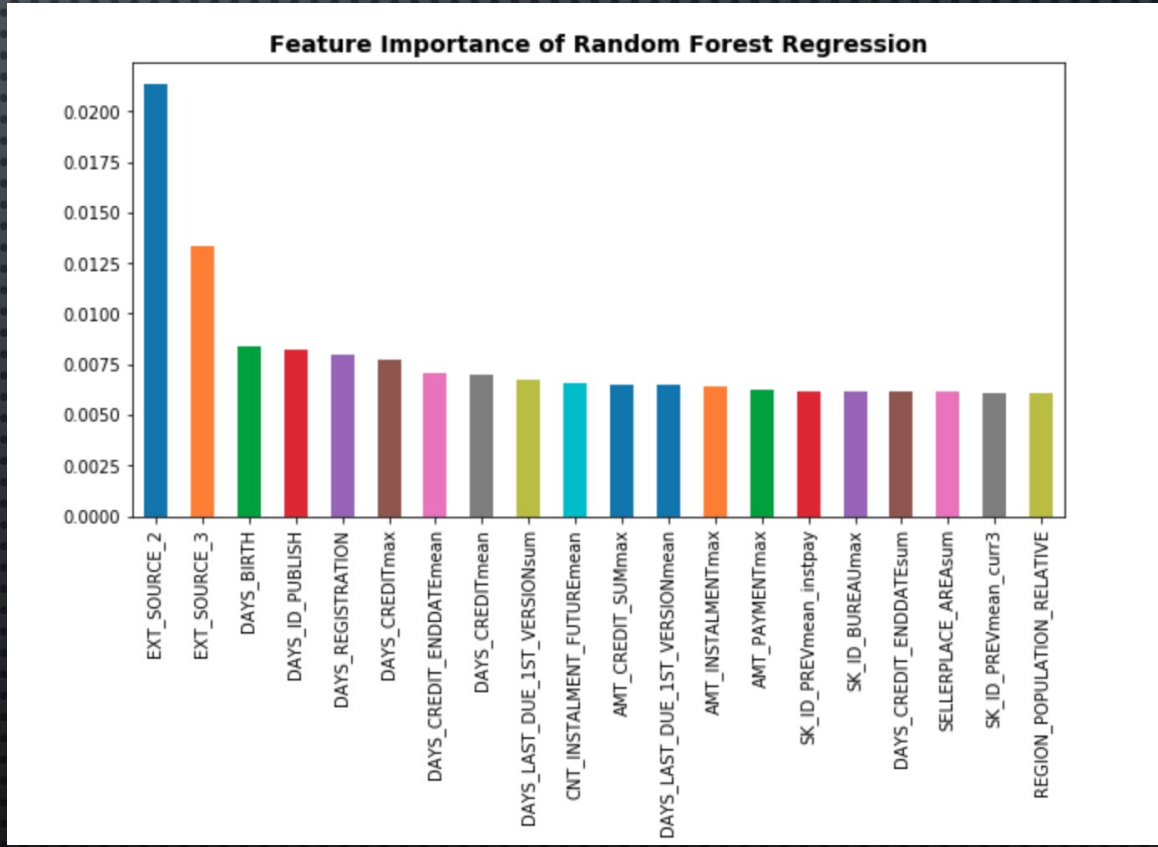
1. Delinquency ↑ as Age ↓
2. Delinquency ↑ as ID_PUBLISH ↑
3. Delinquency ↑ as DAYS_LAST_PHONE_CHANGE ↑

Random Forest Feature Importance

- Top 20 Features explains about 16% of the variance

- Important Features:
  - Ext_Source 2 & 3: Normalized score of external data
  - Days_Birth: Age of Applicant
  - Days_ID_Publish: Number of Days before the application did the applicant change their identity

- Overfitted model which did not generalize well to the dataset
-
- Random Forest AUC Score: .58
- Random Forest F1 Score: .11



Future Work & Conclusion

- **Missing Values:** Large amounts of missing values resulted due to the sparsity of the dataset.
- **Outliers:** Work on eliminating outliers
- **Feature Elimination:** Reduce features of data set to decrease overfitting
- **Machine Learning:** Use additional ML classifiers like LightGBM
-
- **Conclusion:** Difficult to understand credit information without historical financial information. Current model does not generalize well to model, however future work should improve overall predictive capabilities.