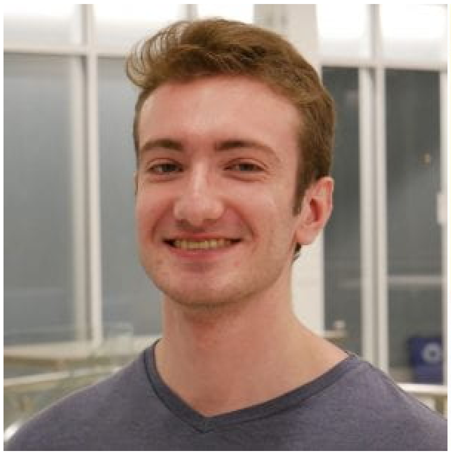
HCDR Team 1 Bears

Zach Sigler
zseliger@gmail.com

Keegan Moore
keegmoor@iu.edu

Rajasimha
ragallam@iu.edu

Jagan Lakku
slakku@iu.edu



Abstract

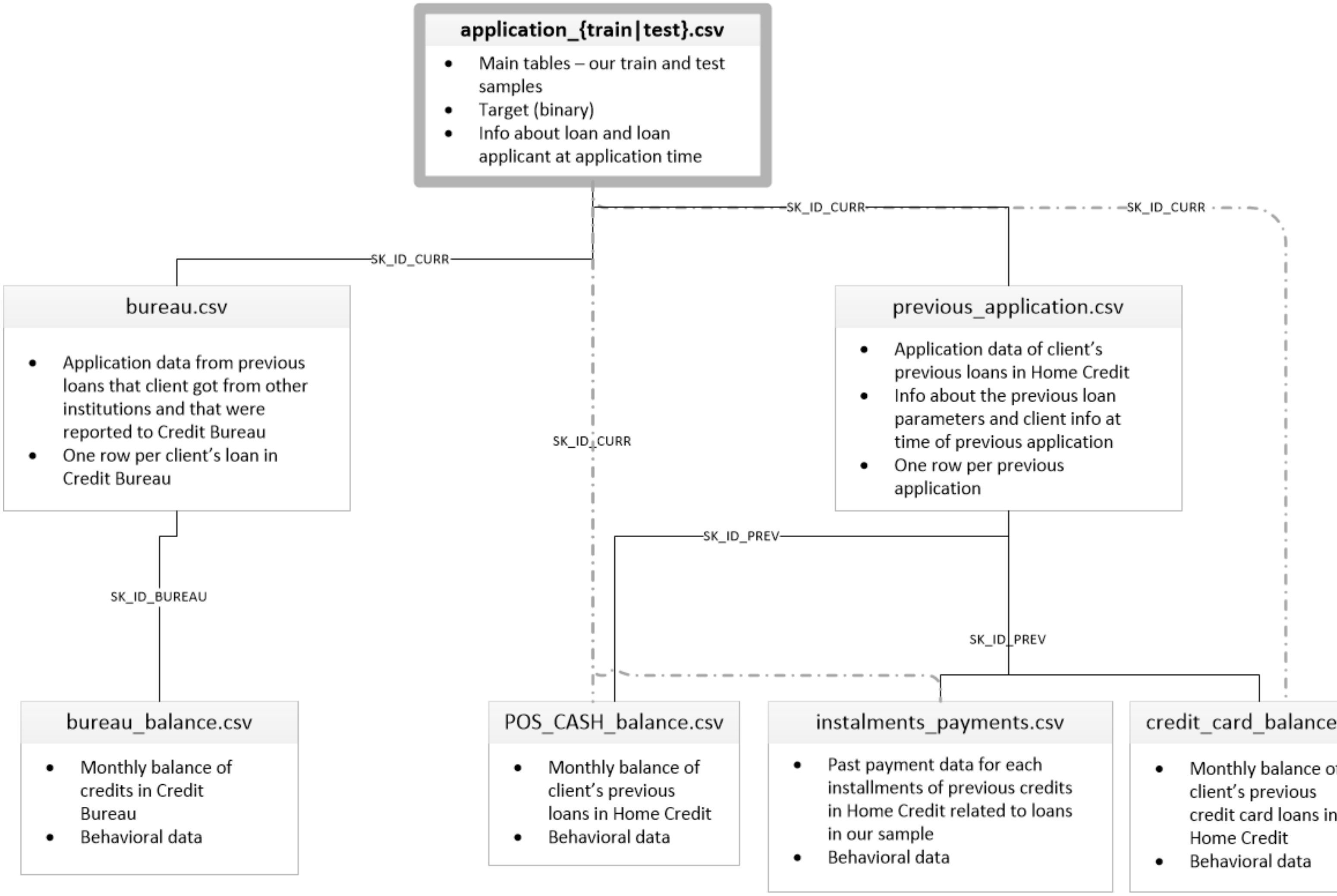
The goal of this project is to use a machine learning model that can predict whether or not someone would be able to repay a loan.

In Phase 1, we conducted EDA on not only the primary application data, but on every secondary database as well. We also create pipelines that use Scalars, Imputers, OHEs, and finally LinearRegression on both the primary and combined datasets. We used the features that had either high correlations with the target feature, or looked interesting when graphed during EDA.

We found that the vast majority of features available were not very useful, and so discarded many. The roc_auc score for the primary dataset alone was .7342, which would place us 5773th out of 7176 entries on the public leaderboards. The score for the combined data was .7401, which would place us at 5563th. Our future plans for increasing our score include testing multiple different learning algorithms beyond LinearRegression, which was chosen as a baseline due to its speed and simplicity.

Project Description

Our data is split up among several different CSV files and looks like this:



application_{train|test}.csv

- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time

bureau.csv

- Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
- One row per client's loan in Credit Bureau

previous_application.csv

- Application data of client's previous loans in Home Credit
- Info about the previous loan parameters and client info at time of previous application
- One row per previous application

bureau_balance.csv

- Monthly balance of credits in Credit Bureau
- Behavioral data

POS_CASH_balance.csv

- Monthly balance of client's previous loans in Home Credit
- Behavioral data

instalments_payments.csv

- Past payment data for each installments of previous credits in Home Credit related to loans in our sample
- Behavioral data

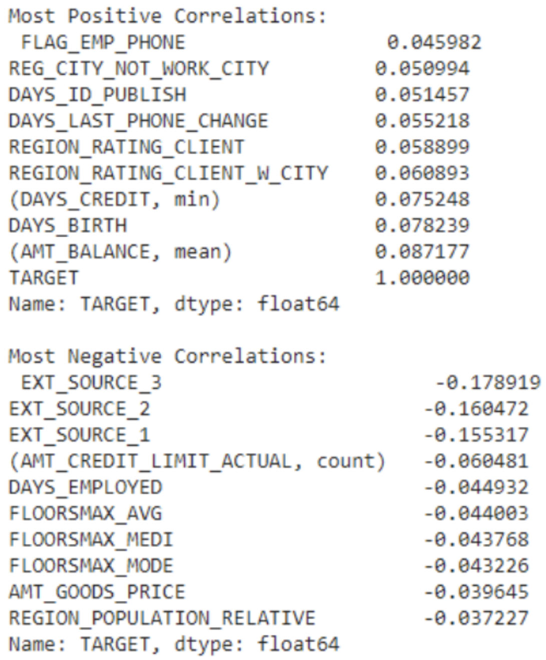
credit_card_balance

- Monthly balance o
client's previous credit card loans in Home Credit
- Behavioral data

Application{train|test}.csv contains static data for all applications and one row represents one loan. SK_ID_CURR represents the ID for that row, which can be tied in with the other datasets for this problem. The other datasets can contain 0, 1, or more documents that correspond to the SK_ID_CURR, which will have to be accounted for in our feature engineering.

Our task was to combine the secondary and primary datasets and train a baseline model on that dataset. We figure out what features to use in training the model through EDA, using correlation and visualization to choose the best features.

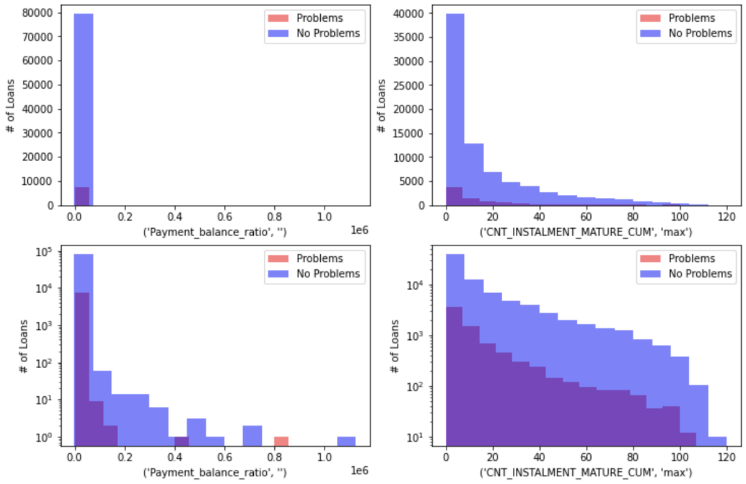
EDA and EDA Visualization



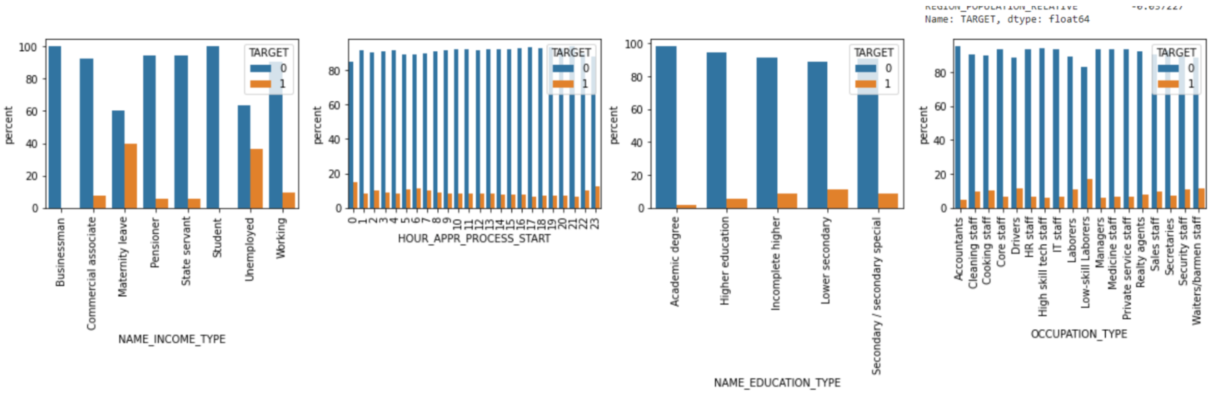
```
Most Positive Correlations:
 FLAG_EMP_PHONE               0.045982
REG_CITY_NOT_WORK_CITY       0.050994
DAYS_ID_PUBLISH              0.051457
DAYS_LAST_PHONE_CHANGE       0.055218
REGION_RATING_CLIENT         0.058899
REGION_RATING_CLIENT_W_CITY  0.060893
(DAYS_CREDIT, min)           0.075248
DAYS_BIRTH                   0.078239
(AMT_BALANCE, mean)          0.087177
TARGET                       1.000000
Name: TARGET, dtype: float64

Most Negative Correlations:
 EXT_SOURCE_3                       -0.178919
EXT_SOURCE_2                       -0.160472
EXT_SOURCE_1                       -0.155317
(AMT_CREDIT_LIMIT_ACTUAL, count)   -0.060481
DAYS_EMPLOYED                      -0.044932
FLOORSMAX_AVG                      -0.044003
FLOORSMAX_MEDI                     -0.043768
FLOORSMAX_MODE                     -0.043226
AMT_GOODS_PRICE                    -0.039645
REGION_POPULATION_RELATIVE         -0.037227
Name: TARGET, dtype: float64
```

The above chart shows the most correlated features from all datasets, which we used in training. Any feature that is a pair came from a secondary dataset.

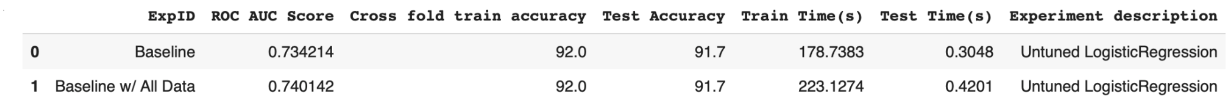
We also tried to create more features from our datasets, but the ones we created did not perform well, and were not used. The below chart shows one such feature, which is the ratio of a credit payment over the total balance on that credit.



The next set of charts shows the most useful categorical features from the application dataset.


```
REGION_POPULATION_RELATIVE          -0.037227
Name: TARGET, dtype: float64
```

Results



	ExpID	ROC AUC Score	Cross fold train accuracy	Test Accuracy	Train Time(s)	Test Time(s)	Experiment description
0	Baseline	0.734214	92.0	91.7	178.7383	0.3048	Untuned LogisticRegression
1	Baseline w/ All Data	0.740142	92.0	91.7	223.1274	0.4201	Untuned LogisticRegression

We had 2 baseline models, one with only the application dataset, and one with all of the datasets.The ROC score of the first was 0.734214, and the second was 0.740142. So, there was a slight improvement from adding the new features and data, though it wasn't a lot. The train and test accuracy were 92 and 91.7 percent respectively, for both models. Clearly more feature engineering can be done, and we may want to refine our feature selection process.

Conclusion

Our goal is to build a machine learning algorithm that can predict whether or not someone will have trouble paying back a loan. This is important because an accurate prediction will be able to help those that are on the edge of getting a loan, but need one desperately. Well thought-out algorithms will be able to accurately predict such things, especially with the volume of data provided. So far, we have only used a baseline regression model and already we are seeing accuracy metrics in the around 92%.We will take this further with more feature engineering, and more clever models than linear regression.