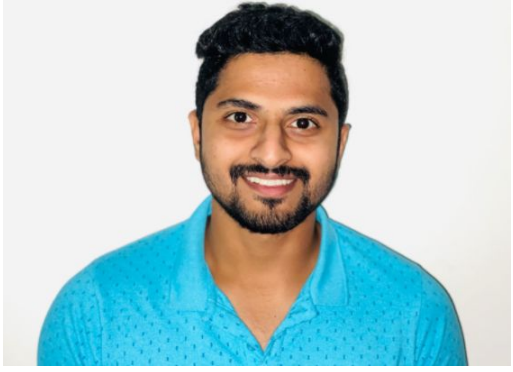
Applied Machine Learning Phase 1

Group 1:
Zack Seliger, Keegan Moore, Jagan Lakku, Raja



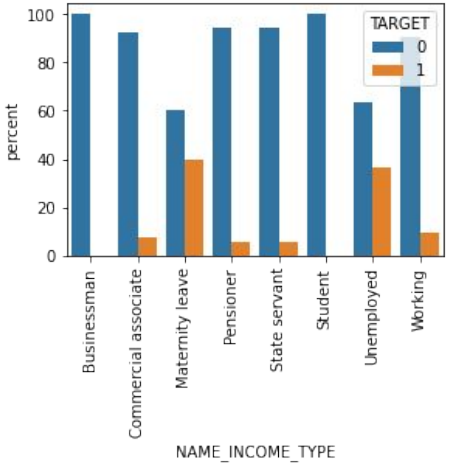


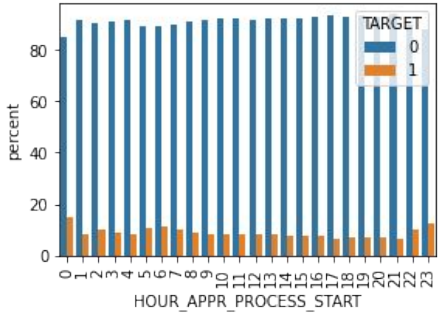

EDA + baselines

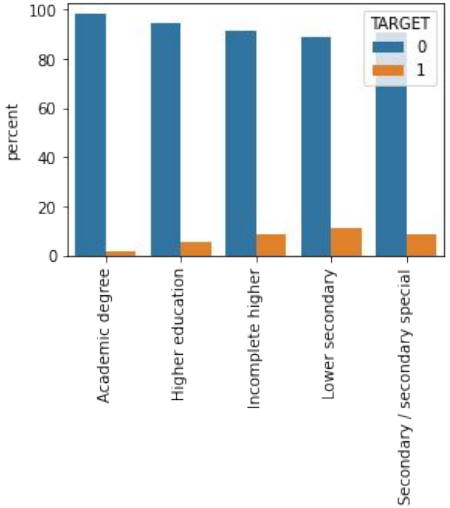
- Found interesting application and secondary features
- 0.74 ROC score using Logistic Regression

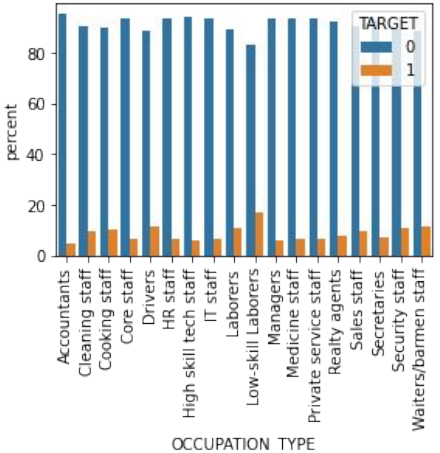
```
Most Positive Correlations:
 FLAG_EMP_PHONE                 0.045982
REG_CITY_NOT_WORK_CITY          0.050994
DAYS_ID_PUBLISH                 0.051457
DAYS_LAST_PHONE_CHANGE          0.055218
REGION_RATING_CLIENT            0.058899
REGION_RATING_CLIENT_W_CITY     0.060893
(DAYS_CREDIT, min)              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
```



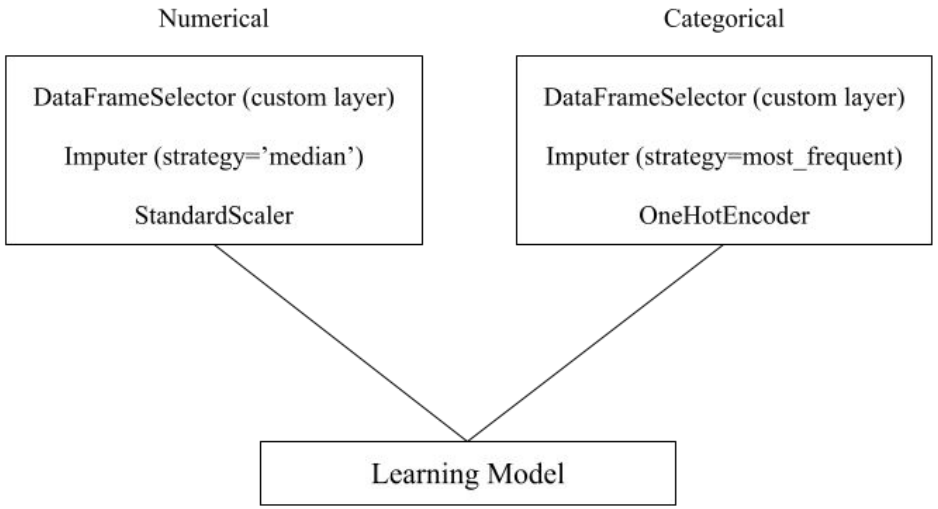






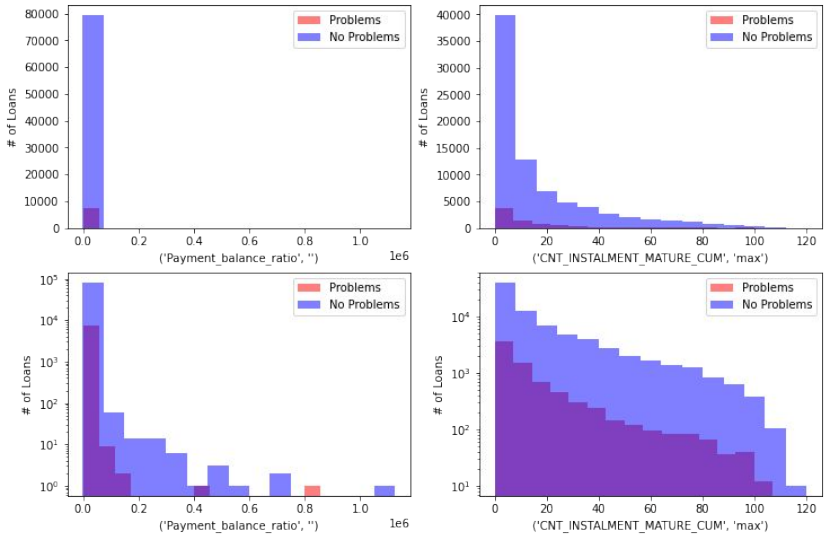
Pipeline

- Our Pipeline follows this model



Feature Engineering

- Experimented aggregating and creating new features
  - Looking at aggregations differently helped a lot
  - New features performed terribly
- Moving forward, need to go deeper into our datasets
- Want to add a lot more data



Summary

- Completed baseline models
- Did basic feature engineering
- Want to add more features to model
- Have to figure out the deeper connections between data sources
- Google Colab is finicky with too much data