Home Credit Default Risk

Team Members:

- Manas Rai
- Aadithya Anandaraj
- Vishal Ramachandran
- Derrick Hung
- Ian Hatfield

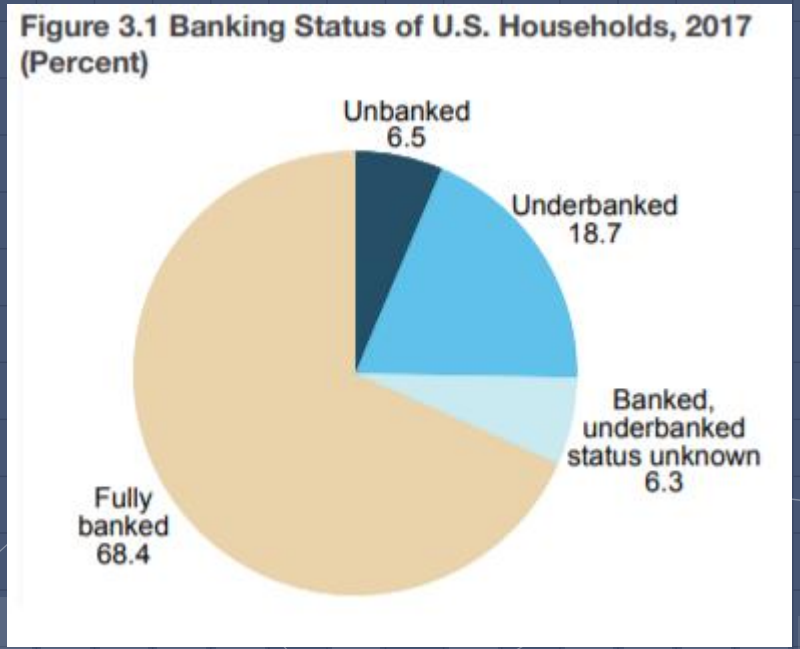
USA Unbanked Population

Unbanked and Underbanked

- Do not have bank account or lack access
- Reliant on alternative financial services
- The Federal Reserve estimated there are 55 million unbanked or underbanked adult

Home Credit

- Providing consumer lending to client without credit history
- Our Task is to predict the default risk of the clients based on the data



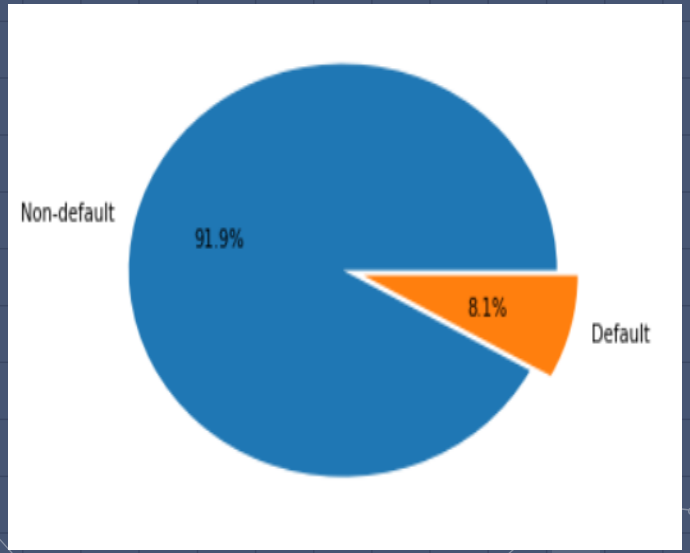
Agenda

1. Data Description
2. Data Treatment
3. Exploratory Data Analysis
4. Model Building
5. Inference

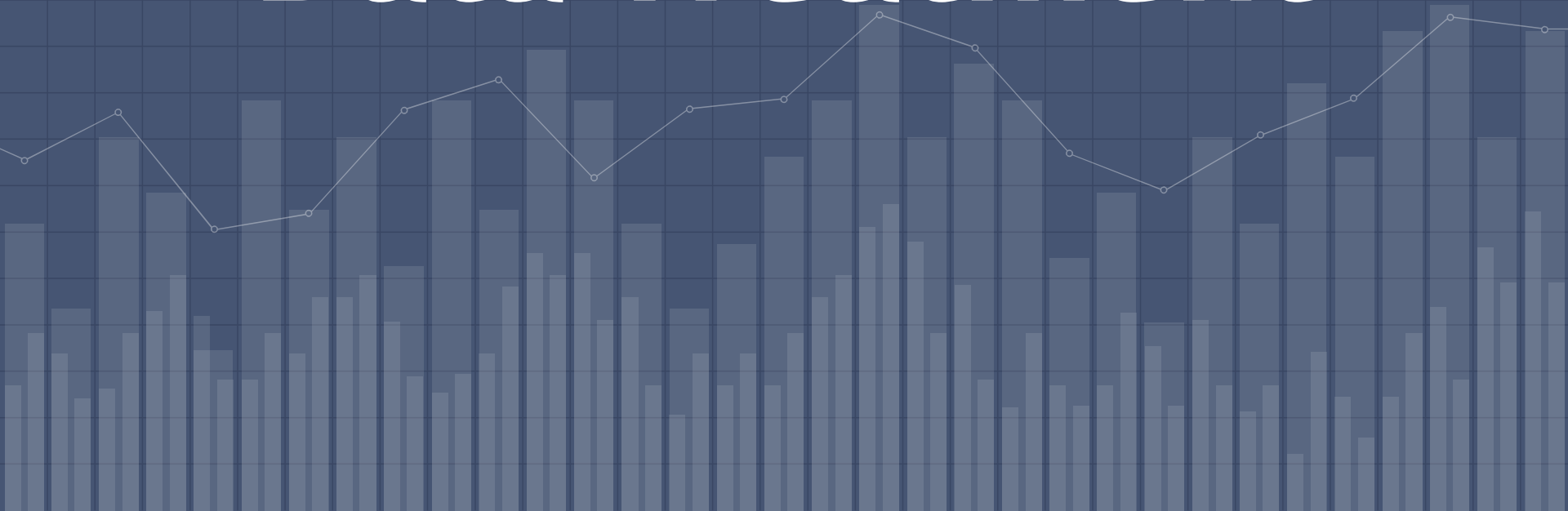
Data Description

Data Description

- **Home Credit default risk dataset** from Kaggle- Info about loan and loan applicants
- **Train set :** 307,511 applicants and 122 features from the application
- **Test :** 48,744 applicants
- **Target :** Binary (defaulter vs non-defaulter)
- **Imbalanced data :** 91.9% non-defaulters



Data Treatment



Data Treatment

Null Value Treatment

- 23 columns with more than 70% nulls were dropped
- Analyzed variables with lesser null values and imputed with mean/median/mode
- Used regression to impute 30 continuous variables with more than 20% nulls

Resampling

Created different sets to handle data imbalance

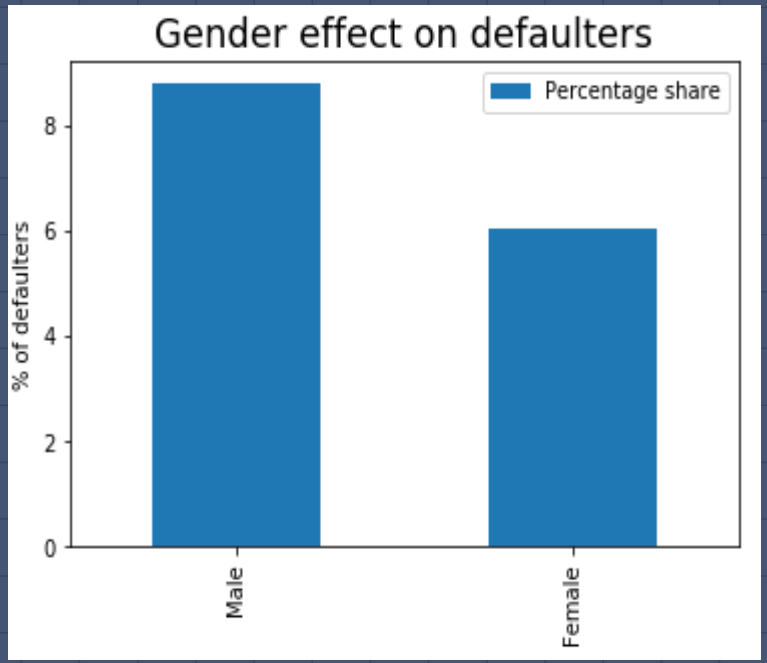
- Random undersampling
- Random oversampling
- Synthetic Minority Over-sampling Technique(SMOTE)

Feature Engineering

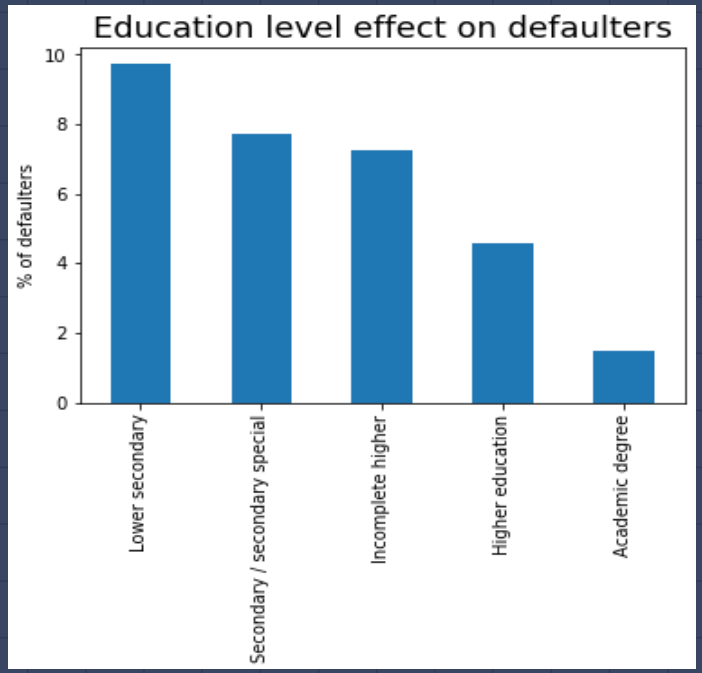
- Created new features by considering interaction between existing features
- Created bins for skewed and sparsely distributed continuous variables like family size and no. of children

Exploratory Data Analysis

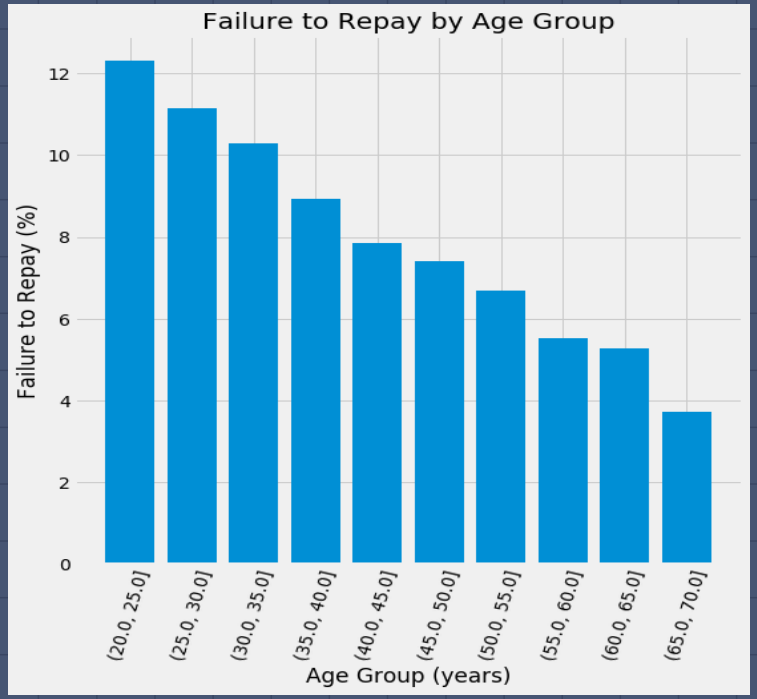
Females pay back more



Learned pay back more



Younger people pay back less



Low-skill labourers pay back less

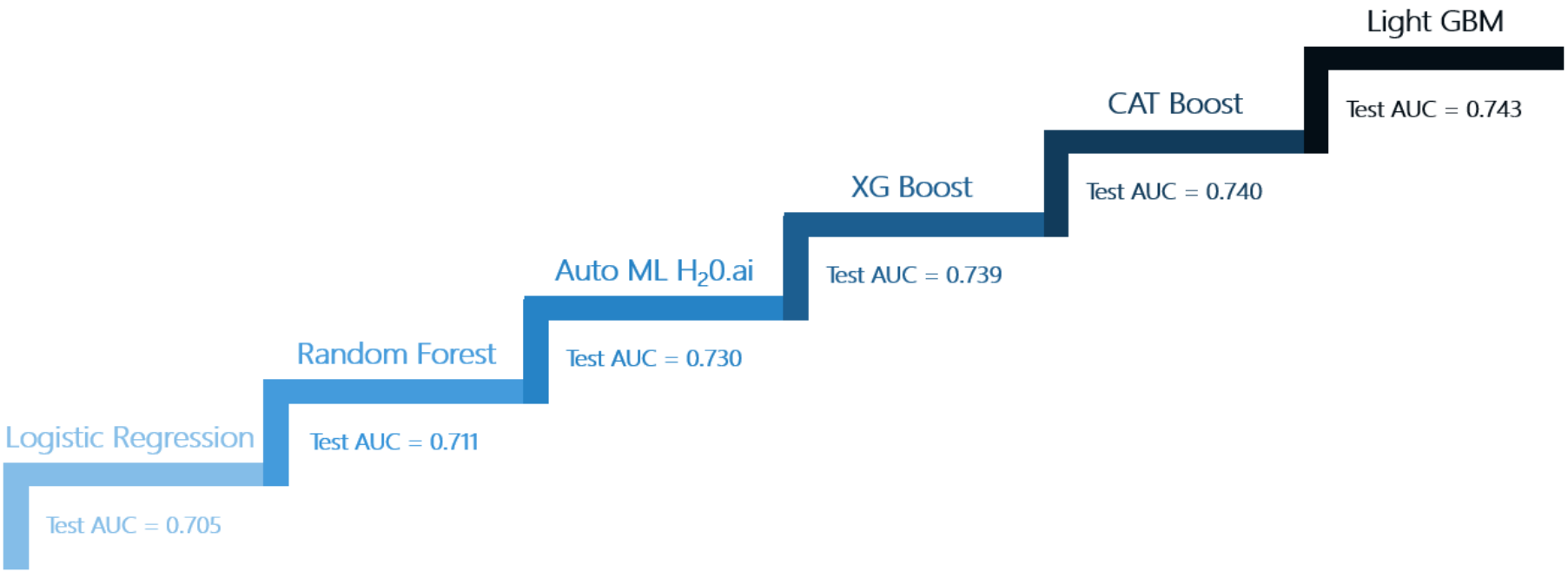
Basically. . EDA says

A middle aged female degree holder who is a manager

A young male with a secondary level education who is a low-skill labourer

Modeling

Models Outcome



Logistic Regression

Unsampled Data

- **Train AUC: 0.742**
- **Test AUC: 0.705**

Resampled Data

- **Train AUC: 0.871**
- **Test AUC: 0.711**

```
Coefficients: (18 not defined because of singularities)
                                Estimate Std. Error    z value Pr(>|z|)
(Intercept)                -4.401e+24  1.665e+17 -2.643e+07   <2e-16 ***
AMT_ANNUITY                 4.592e+15  6.205e+06  7.400e+08   <2e-16 ***
AMT_CREDIT                  1.974e+15  8.845e+06  2.231e+08   <2e-16 ***
AMT_GOODS_PRICE            -2.524e+14  8.755e+06 -2.883e+07   <2e-16 ***
AMT_INCOME_TOTAL            2.246e+15  6.362e+07  3.531e+07   <2e-16 ***
AMT_REQ_CREDIT_BUREAU_DAY  -1.275e+14  1.117e+07 -1.142e+07   <2e-16 ***
AMT_REQ_CREDIT_BUREAU_HOUR  2.802e+14  6.391e+06  4.384e+07   <2e-16 ***
AMT_REQ_CREDIT_BUREAU_MON  -2.260e+14  3.852e+06 -5.866e+07   <2e-16 ***
AMT_REQ_CREDIT_BUREAU_QRT  -1.286e+15  4.266e+07 -3.015e+07   <2e-16 ***
AMT_REQ_CREDIT_BUREAU_WEEK  5.614e+14  5.216e+06  1.076e+08   <2e-16 ***
AMT_REQ_CREDIT_BUREAU_YEAR -2.373e+15  1.833e+06 -1.295e+09   <2e-16 ***
APARTMENTS_AVG              5.110e+14  2.061e+07  2.480e+07   <2e-16 ***
APARTMENTS_MEDI            -8.672e+14  2.281e+07 -3.802e+07   <2e-16 ***
APARTMENTS_MODE            -1.652e+14  1.328e+07 -1.244e+07   <2e-16 ***
CNT_CHILDREN               -7.040e+17  1.298e+10 -5.424e+07   <2e-16 ***
CNT_FAM_MEMBERS             7.074e+17  1.298e+10  5.451e+07   <2e-16 ***
DAYS_BIRTH                  5.143e+14  1.015e+06  5.066e+08   <2e-16 ***
DAYS_EMPLOYED              -5.571e+14  1.167e+06 -4.775e+08   <2e-16 ***
DAYS_ID_PUBLISH            -3.464e+14  6.203e+05 -5.584e+08   <2e-16 ***
DAYS_LAST_PHONE_CHANGE      4.430e+13  6.754e+05  6.559e+07   <2e-16 ***
DAYS_REGISTRATION          -1.692e+14  9.240e+05 -1.831e+08   <2e-16 ***
DEF_30_CNT_SOCIAL_CIRCLE   -3.001e+15  1.861e+07 -1.613e+08   <2e-16 ***
DEF_60_CNT_SOCIAL_CIRCLE    9.651e+14  1.581e+07  6.106e+07   <2e-16 ***
ELEVATORS_AVG              -1.563e+15  2.044e+07 -7.648e+07   <2e-16 ***
ELEVATORS_MEDI              2.049e+15  2.350e+07  8.721e+07   <2e-16 ***
ELEVATORS_MODE             -1.364e+14  1.226e+07 -1.113e+07   <2e-16 ***
ENTRANCES_AVG               1.363e+15  2.716e+07  5.018e+07   <2e-16 ***
ENTRANCES_MEDI             -1.539e+15  2.879e+07 -5.346e+07   <2e-16 ***
ENTRANCES_MODE              3.841e+14  1.213e+07  3.168e+07   <2e-16 ***
EXT_SOURCE_1               -1.696e+15  1.166e+06 -1.454e+09   <2e-16 ***
EXT_SOURCE_2                1.386e+14  6.041e+05  2.295e+08   <2e-16 ***
EXT_SOURCE_3                2.622e+14  6.485e+05  4.043e+08   <2e-16 ***
FLAG_CONT_MOBILE            2.092e+14  2.899e+06  7.215e+07   <2e-16 ***
FLAG_DOCUMENT_10           -2.200e+15  2.537e+07 -8.670e+07   <2e-16 ***
```

Random Forest

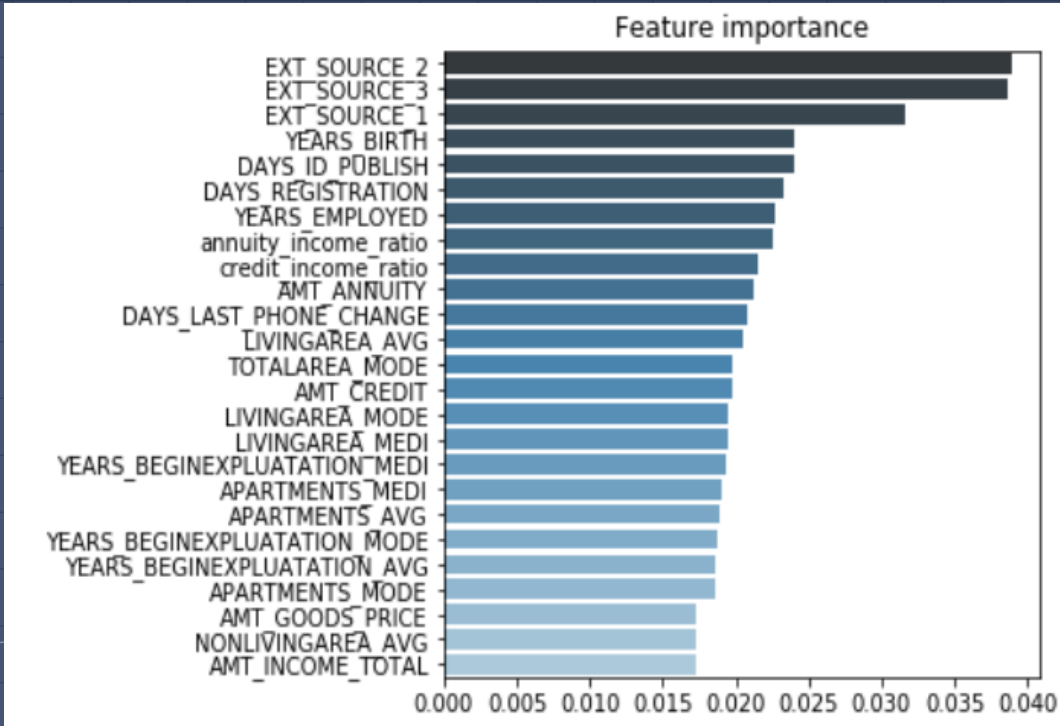
- 1000 trees

Unsampled Data

- **Test AUC: 0.711**

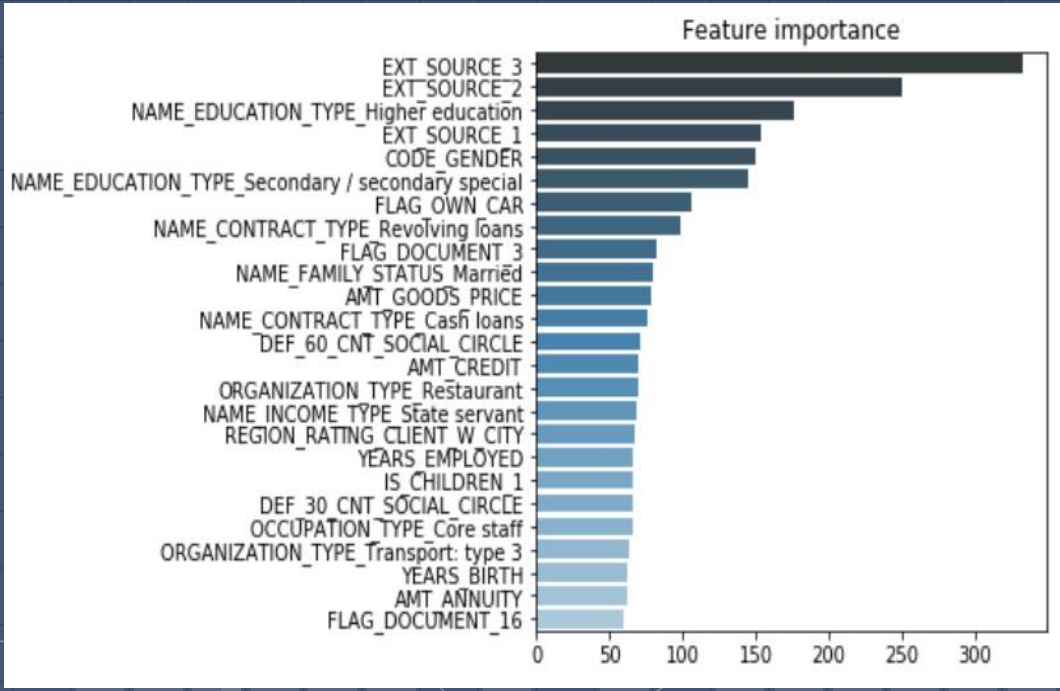
Resampled Data

- **Test AUC: 0.711**



XGBoost

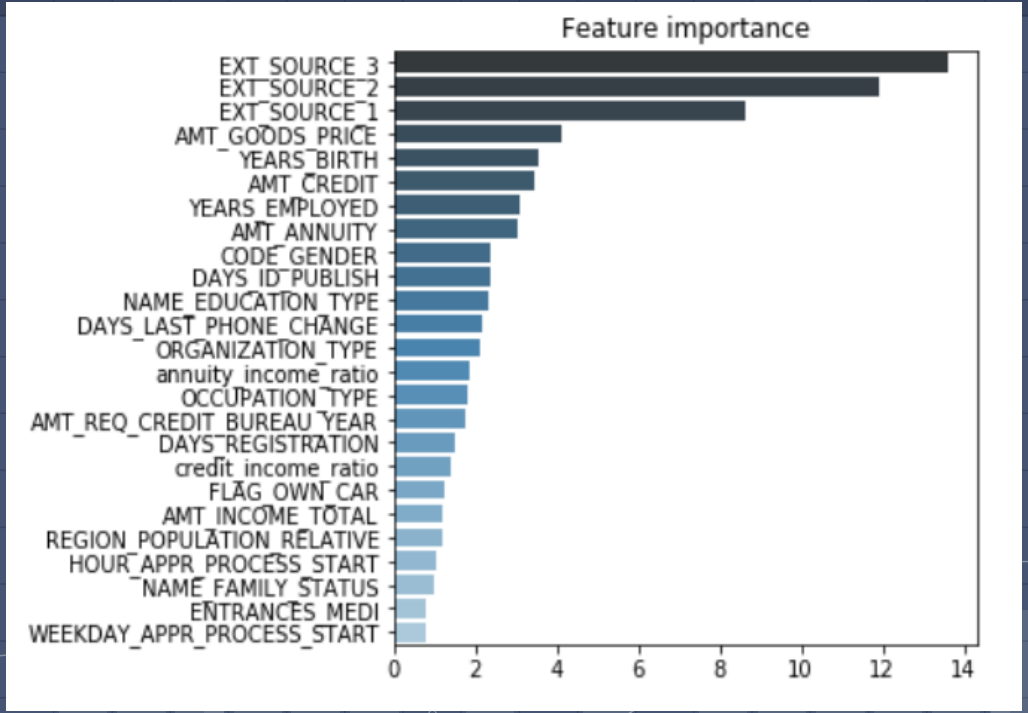
- 309 boosting trees
- Tuned Parameters
  - Learning rate: 0.03
  - Subsample: 0.8
  - Colsample: 0.8
  - Max depth: 7
- **Train AUC: 0.758**
- **Test AUC: 0.739**



CatBoost

- 2004 boosting trees
- Learning rate: 0.03
- Subsample: 0.7
- Colsample: 0.7
- Max depth: 7

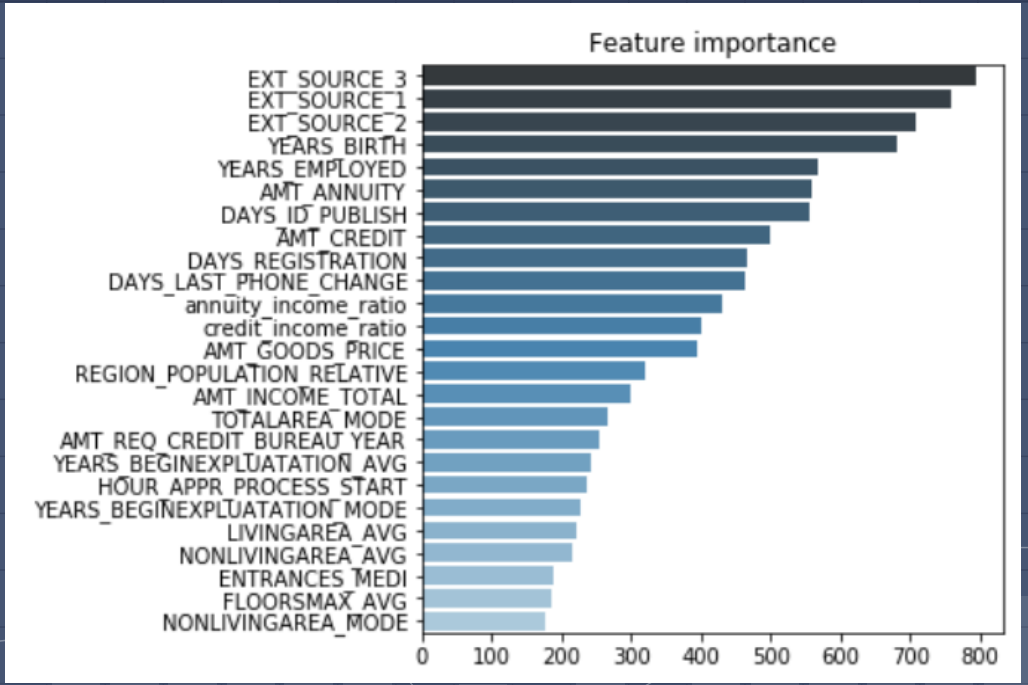
- **Train AUC: 0.763**
- **Test AUC: 0.740**



Light GBM

- 394 boosting trees
- Learning rate: 0.05
- Num_leaves: 31

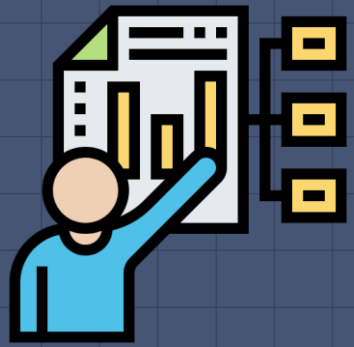
- **Train AUC: 0.758**
- **Test AUC: 0.743**



Inference

Key takeaways

- Low external score is a strong indicator for default
- Customer Age and duration of employment influence the default rate



- Resampling data led to overfitting
- Boosting works best but never underestimate Logistic Regression

Males should be scrutinized more

Thank you!