Lukas Sykora

# Get loans with Home Credit

Kaggle competition
(https://www.kaggle.com/c/home-credit-default-risk)

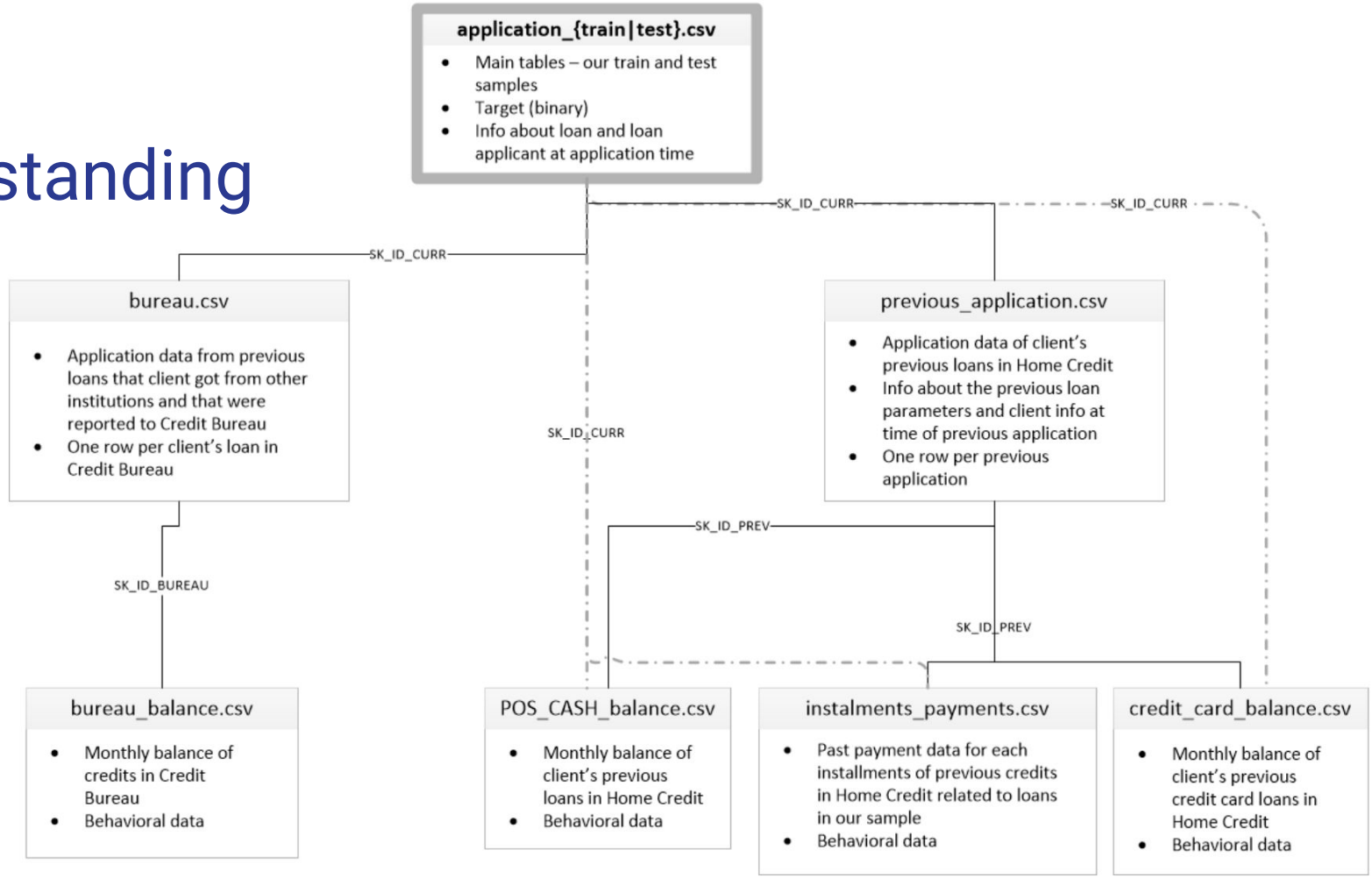
# Workflow


# Business Understanding

**Target**

Ensure that clients capable of repayment are not rejected and that loans are given.

# Data Understanding



# Data Preparation

- Everything to one table.
- Summary rows (GROUP BY) for related tables.
- Aggregate functions (COUNT, MAX, MIN, AVG).

Training data frame:

```
Rows:307511
Cols:613
```

# Modeling

## Gradient Boosting

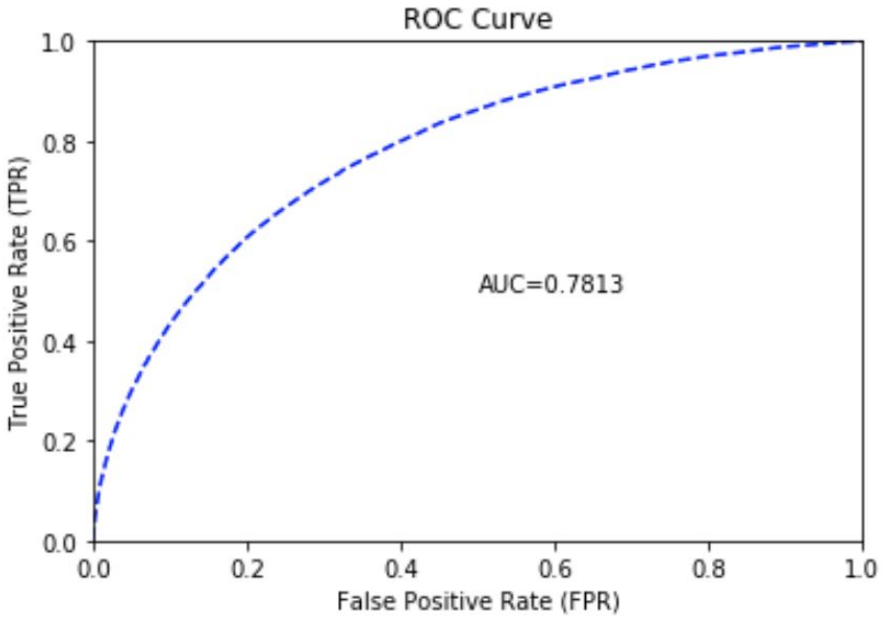
Gradient boosting is a machine learning technique, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

## Auto ML

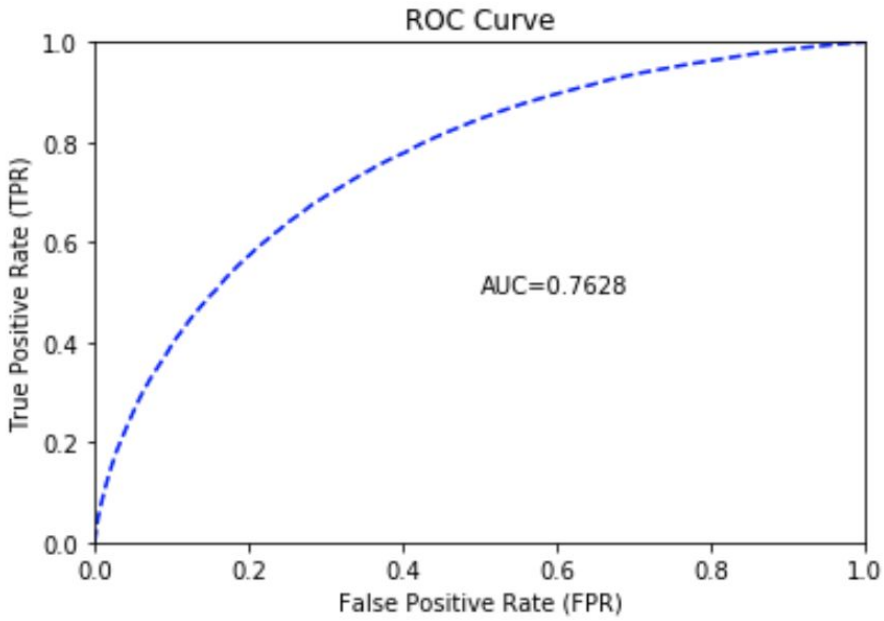
H2O's AutoML can be used for automating the machine learning workflow, which includes automatic training and tuning of many models.

# Evaluation - Gradient Boosting - AUC

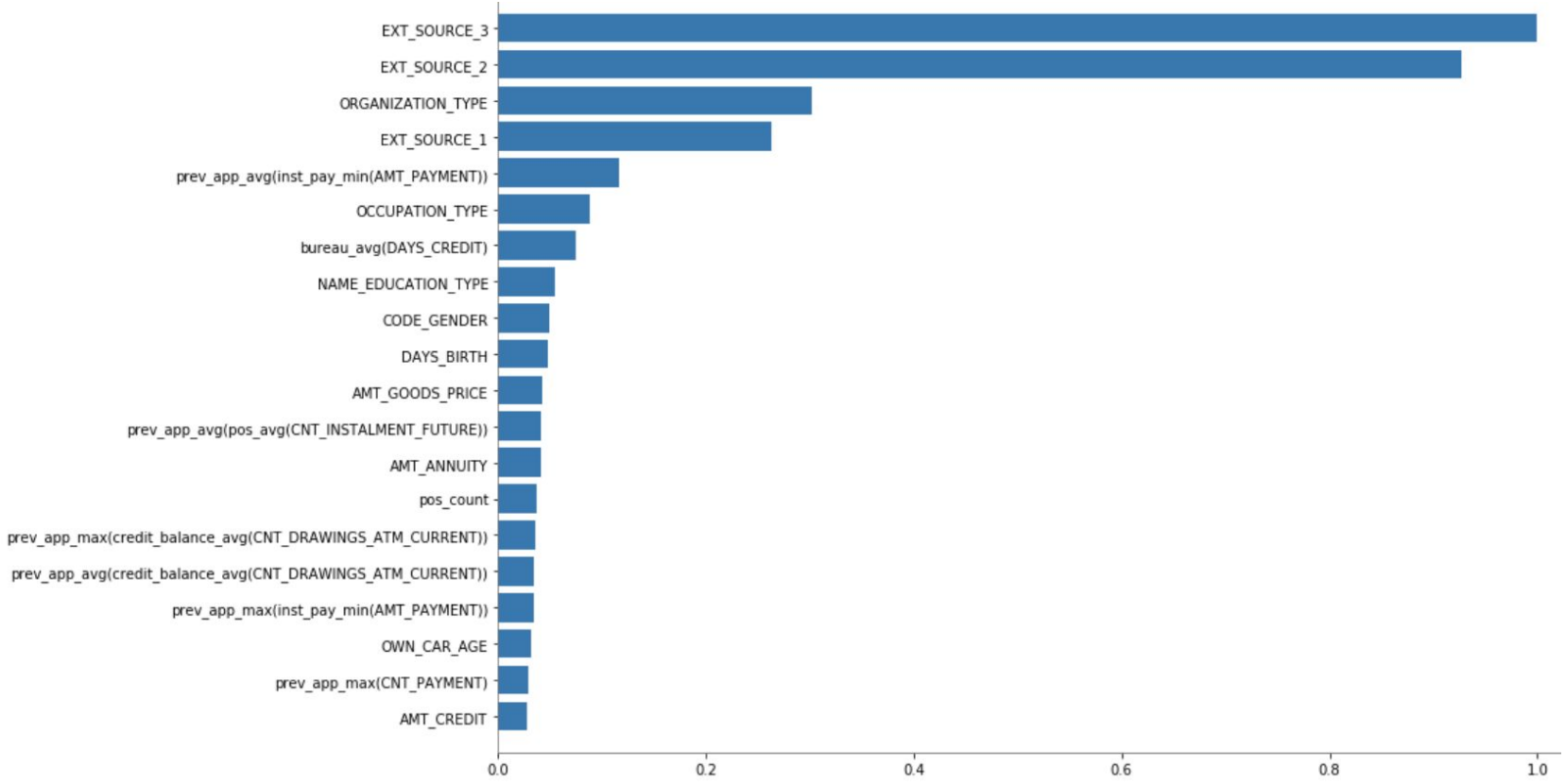
Training Data




X-Val



# Evaluation - Gradient Boosting - Variable Import.



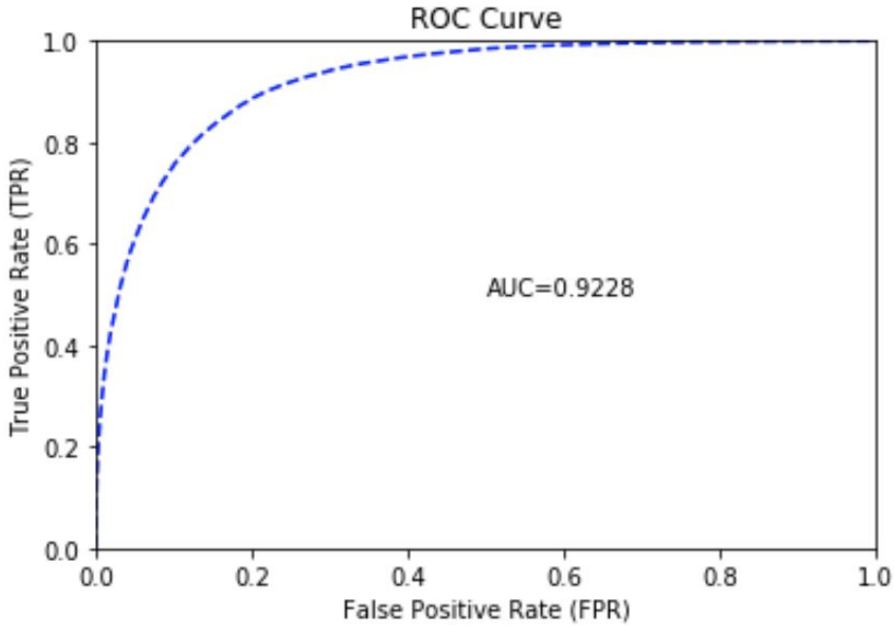
# Evaluation - Auto ML - Leaderboard

| model_id | auc | logloss | aucpr | mean_per_class_error | rmse | mse |
|---|---|---|---|---|---|---|
| StackedEnsemble_AllModels_AutoML_20200415_174004 | 0.781499 | 0.244769 | 0.274235 | 0.331792 | 0.259405 | 0.0672908 |
| StackedEnsemble_BestOfFamily_AutoML_20200415_174004 | 0.780312 | 0.245245 | 0.271549 | 0.329684 | 0.259658 | 0.0674224 |
| XGBoost_3_AutoML_20200415_174004 | 0.774454 | 0.241502 | 0.262289 | 0.337221 | 0.258688 | 0.0669193 |
| GBM_2_AutoML_20200415_174004 | 0.773784 | 0.242093 | 0.260099 | 0.328162 | 0.258919 | 0.0670391 |
| GBM_1_AutoML_20200415_174004 | 0.773303 | 0.242656 | 0.255131 | 0.343667 | 0.259274 | 0.0672231 |
| GLM_1_AutoML_20200415_174004 | 0.767817 | 0.244274 | 0.247271 | 0.339984 | 0.259776 | 0.0674837 |
| XGBoost_1_AutoML_20200415_174004 | 0.747644 | 0.254152 | 0.227684 | 0.353298 | 0.264116 | 0.0697572 |
| XGBoost_2_AutoML_20200415_174004 | 0.730124 | 0.265823 | 0.213109 | 0.357364 | 0.267463 | 0.0715362 |
| DRF_1_AutoML_20200415_174004 | 0.727893 | 0.255445 | 0.201658 | 0.364662 | 0.264182 | 0.0697924 |

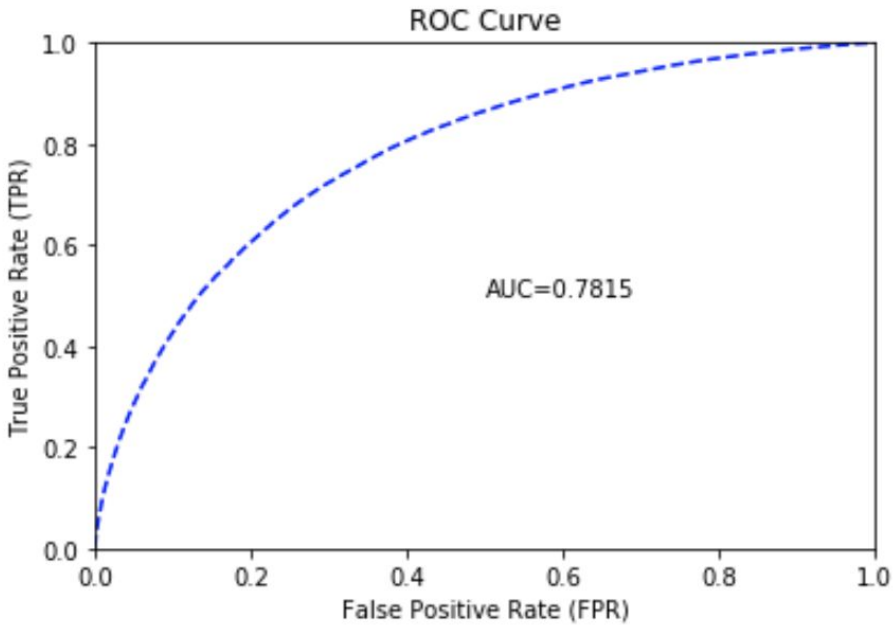
- Stacking uses a “meta learner” (not voting) to combine the predictions of “base learners.”
- XGBoost is an implementation of gradient boosted decision trees.
- Gradient boosting is a machine learning technique for regression and classification problems.
- Generalized Linear Models (GLM), a statistical technique for linear modeling.

# Evaluation - Auto ML (stacked all models) - AUC

`Training Data`




`X-Val`



# Deployment

| Submission and Description | Private Score | Public Score |
|---|---|---|
| model_DRF_1_AutoML_20200415_174004.csv<br>6) DRF | 0.64859 | 0.64591 |
| model_GLM_1_AutoML_20200415_174004.csv<br>5) GLM | 0.64859 | 0.64591 |
| model_GBM_2_AutoML_20200415_174004.csv<br>4) GBM 2 | 0.64859 | 0.64591 |
| model_XGBoost_3_AutoML_20200415_174004.csv<br>3) XGBoost 3 | 0.64859 | 0.64591 |
| model_StackedEnsemble_BestOfFamily_AutoML_20200415_174004.csv<br>2) Stacked - best of family | 0.64859 | 0.64591 |
| model_StackedEnsemble_AllModels_AutoML_20200415_174004.csv<br>1) Stacked - all models | 0.64859 | 0.64591 |
| model_gbm.csv<br>0) GBM | 0.62484 | 0.61508 |