Catboosterization

Project Mcnulty
Andrew Portal

Motivation

General Case: Working with solely or largely categorical data

Business Case: Predicting resource access

Data Overview

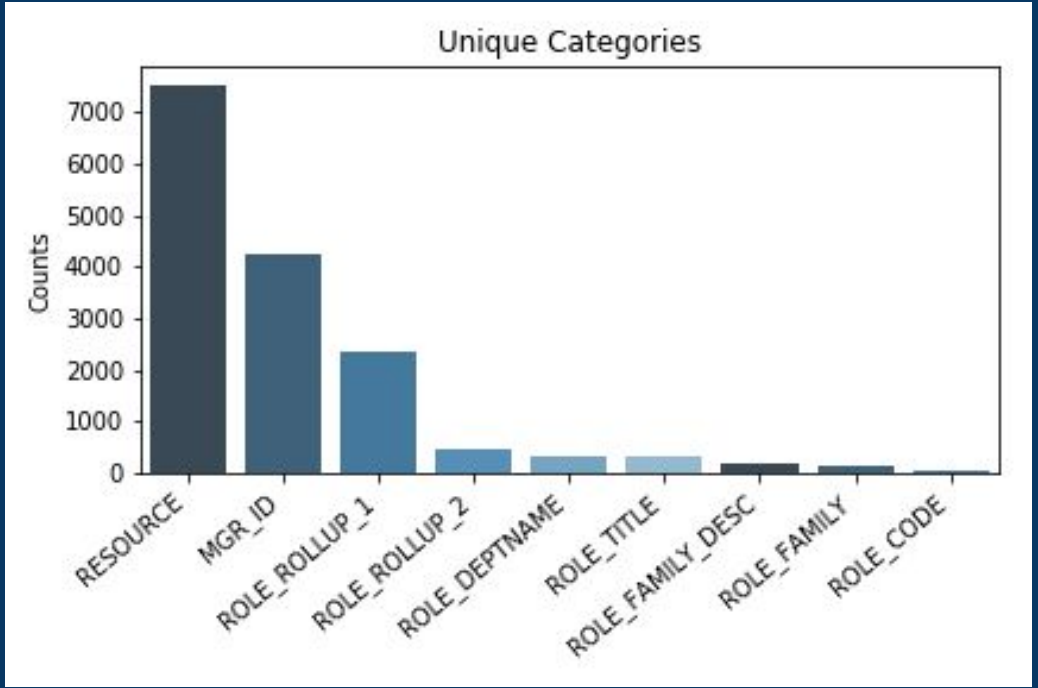
Kaggle Amazon.com - Employee Access Challenge

Target Variable: 0 or 1 (Reject or Approve)

- Targets imbalanced , 1 to 16 ratio

Predictor Variables: 9 features

- Lots of unique categories



Model Selection

Logistic regression: Ok with one-hot and oversampling

KNN: not applicable

SVMs: not applicable

Decision Trees*: poor; sklearn automatically one hot encodes

Catboost: good right out of the box; improves with ensembling

What is Catboost?

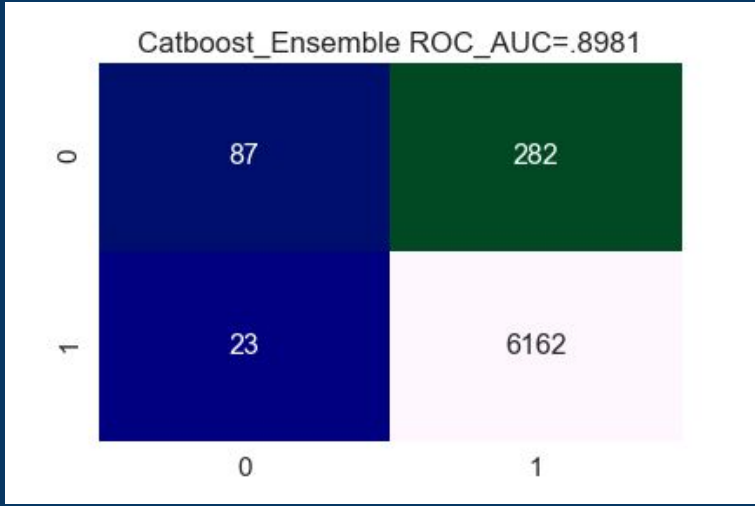
Tree Based

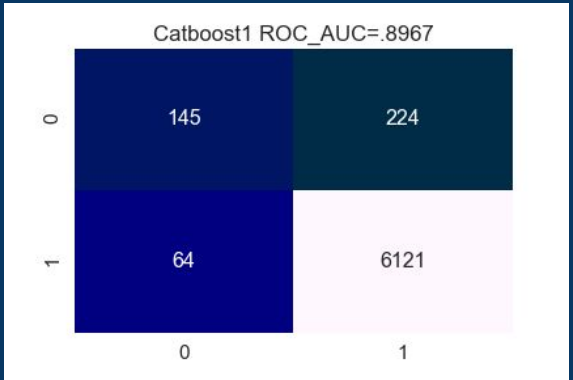
Gradient Boosting + Mapping categorical features to numbers

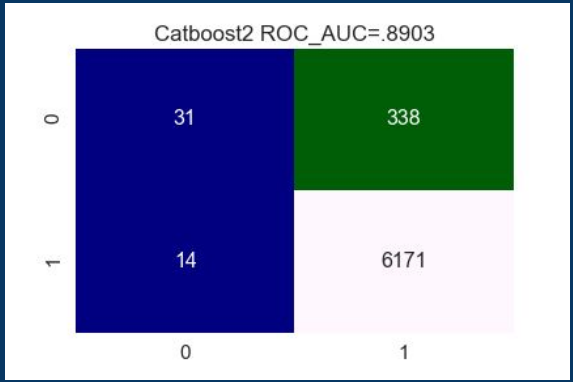
Formula: feature=(countInclass+prior)/(total count+1)

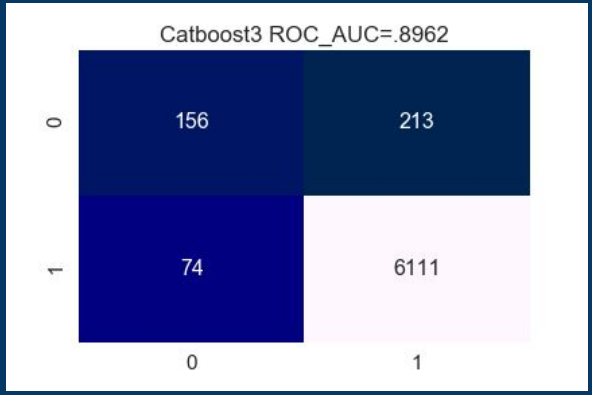
Ensembling

Improvement=.002









App demo

Future Work

More robust parameter search

Model diversity search for ensembling

Try for a tree based anomaly detection algorithm