



EIGENAUTS

Bernard Ong | Emma (Jielei) Zhu | Nanda Rajarathinam | Trinity (Miaozhi) Yu

NYC Data Science Academy Capstone Project
September 2016

CONSUMER CREDIT DEFAULT

KAGGLE CHALLENGE

Improve on the state of the art in credit scoring by predicting the probability that somebody will experience financial distress in the next two years.

Credit scoring algorithms, which make a guess at the probability of default, are the method banks use to determine whether or not a loan should be granted. This challenge seeks to improve on the state of the art in credit scoring, by predicting the probability that somebody will experience financial distress in the next two years.

The goal of this challenge is to build a model that borrowers can use to help make the best financial decisions.

Predict : Serious Delinquency in 2 Years (Default Risk)

Features	Features
Revolving Utilization of Unsecured Lines	Debt Ratio
Age	Monthly Income
Number of Time 30-59 Days Past Due	Number of Open Credit Lines & Loans
Number of Time 60-89 Days Past Due	Number of Real Estate Loans or Lines
Number of Times 90 Days Late	Number of Dependents

GOALS

- **Achieve the highest Area Under the Curve (AUC) Score on the Kaggle Private Leaderboard**
- **Achieve the Top 5% position placement in the Kaggle Private Leaderboard**

DATASET SIZE

- **Training Dataset = 150,000 Observations**
- **Test Dataset = 101,530 Observations**

SWOT ANALYSIS

A SWOT analysis allows our team to organize our thoughts and focus on leveraging our greatest strengths, understand our weaknesses, take advantage of opportunities, and have awareness of any threats to our undertaking.

SWOT allows us to put ourselves on the right track right away, and saves us from a lot of headaches later on.



STRENGTHS

Leveraging what we already know



WEAKNESSES

Areas that we need to improve upon



OPPORTUNITIES

Chance to apply & learn from experience



THREATS

Risks that we need to mitigate and manage

STRENGTHS

- Experience to leverage advanced stacking techniques and Agile
- A synergistic team with very complementary skills and experiences
- Lessons learned from previous Kaggle challenge can be applied
- Experience in the use of Agile process to parallel track work queues

WEAKNESSES

- Extreme time constraints would limit the scope/depth of exploration
- Unfamiliarity with new tools and models will limit firepower
- Learning while executing will slow down the entire process
- Sparse resources on some of the newer technologies employed

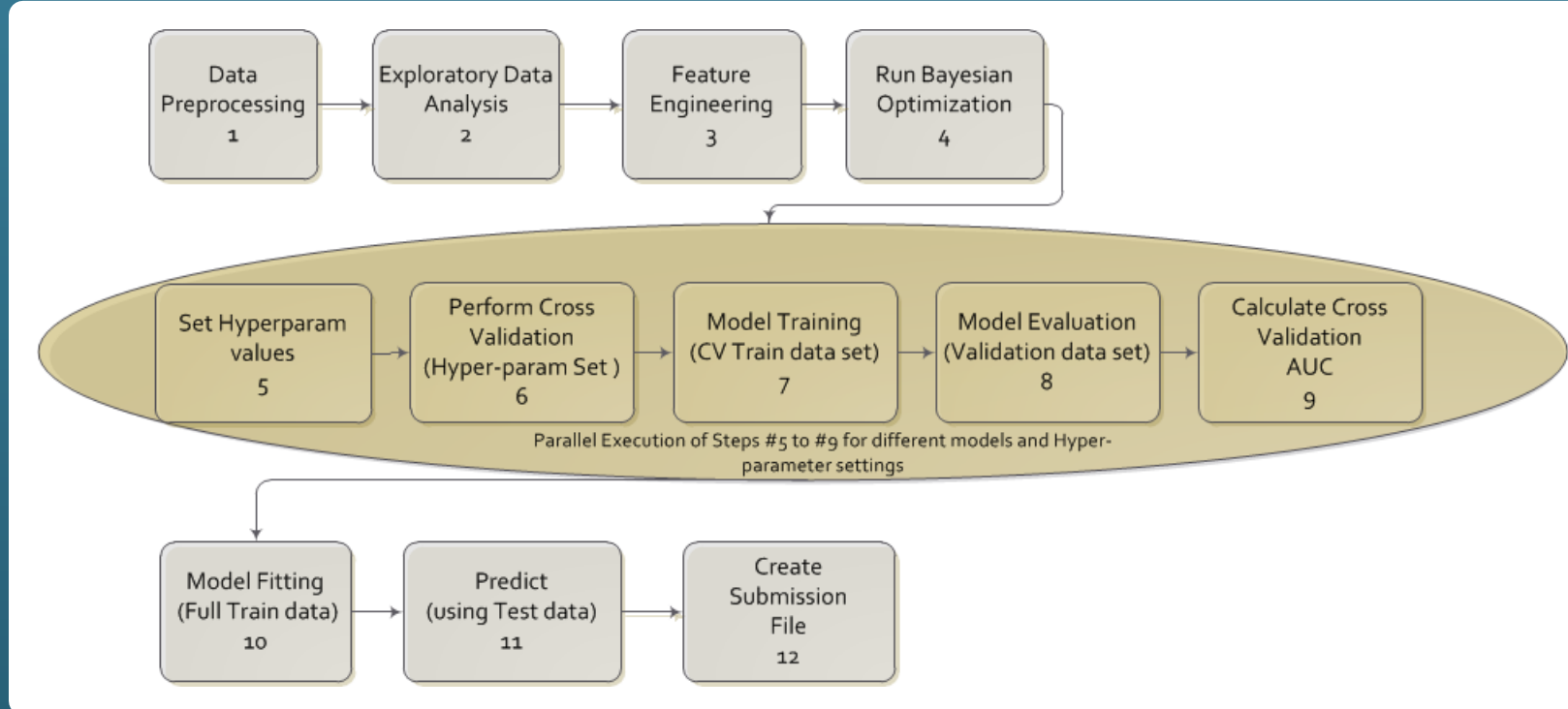
OPPORTUNITIES

- Learn how to strategize, optimize, and fine-tune models, algorithms, and parameters to achieve the best default prediction possible
- Gain real time experience using Bayesian Optimizers
- Explore Deep Learning (Theano/Keras) in predicting default

THREATS

- Small dataset size presents tremendous challenges on generalization that would impact modeling and ultimately, prediction accuracy
- Extremely tight tolerances in the AUC scores in the 10,000th decimal
- Top 5% target goal presents a high risk – no idea how feasible this is

AGILE PROCESS

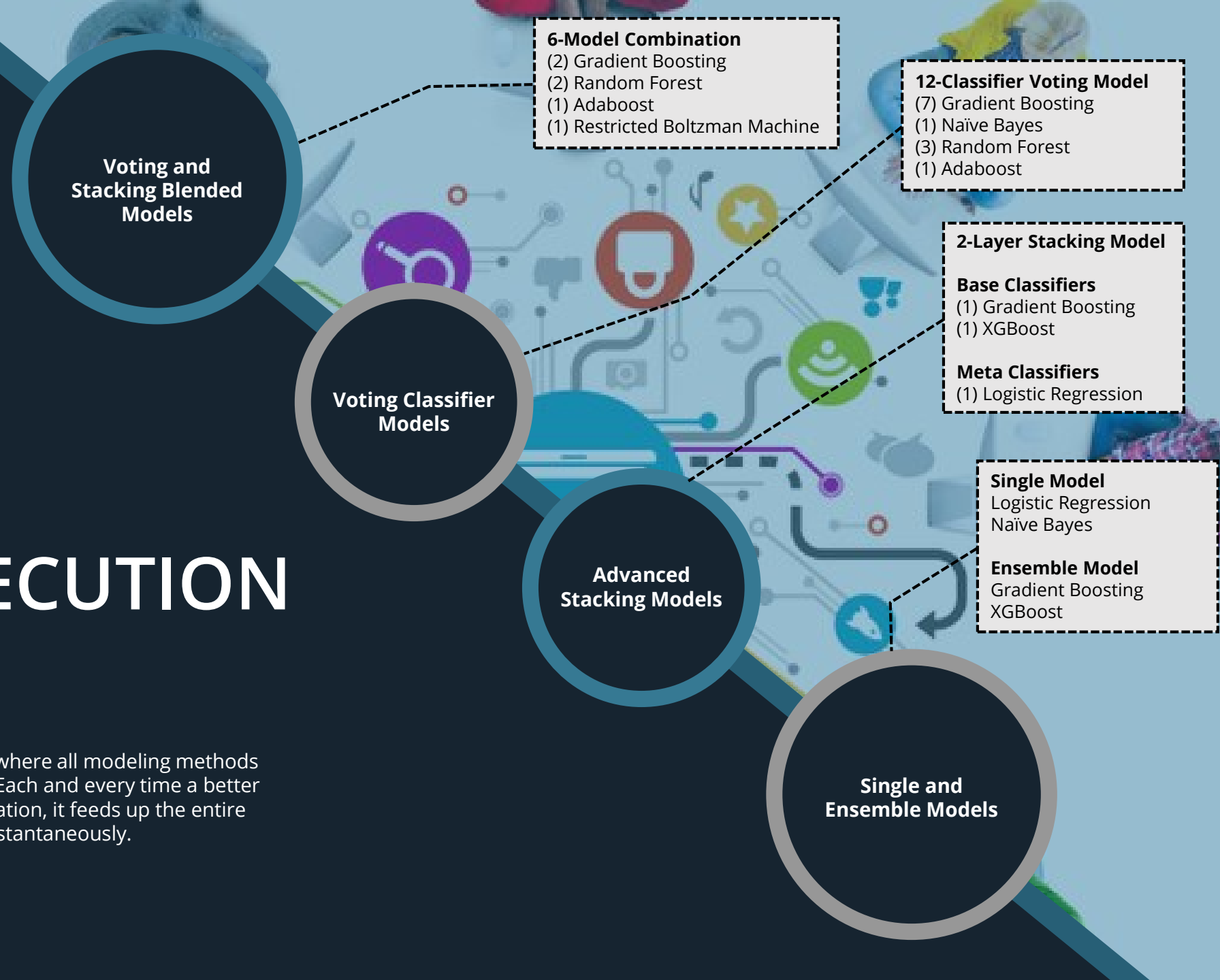


With the constraint of time and resources, it is critical that the right process is utilized to maximize both our throughput and output. We employed an Agile process adapted for Machine Learning that allows us to parallelize our model build, train, validate, test, predict, and scoring cycles quickly in an iterative fashion.

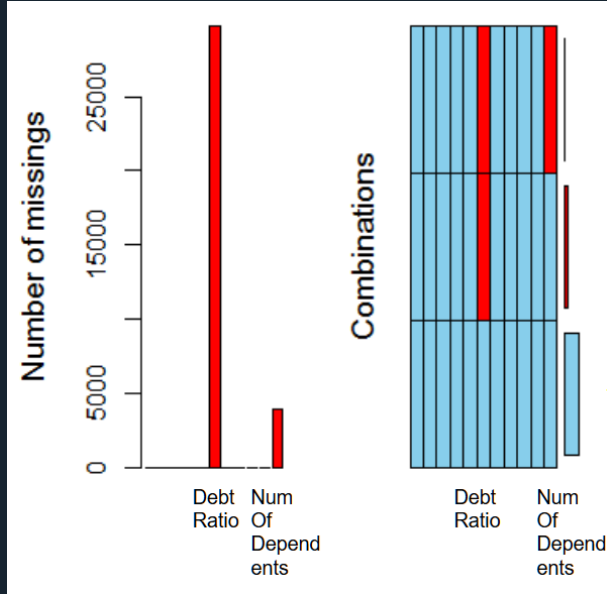
MANAGING COMPLEXITY

MODEL EXECUTION STRATEGY

We employed a parallel tracking process where all modeling methods were being performed at the same time. Each and every time a better setting is found using automated optimization, it feeds up the entire process cycle and synergies are gained instantaneously.



FEATURE ANALYSIS

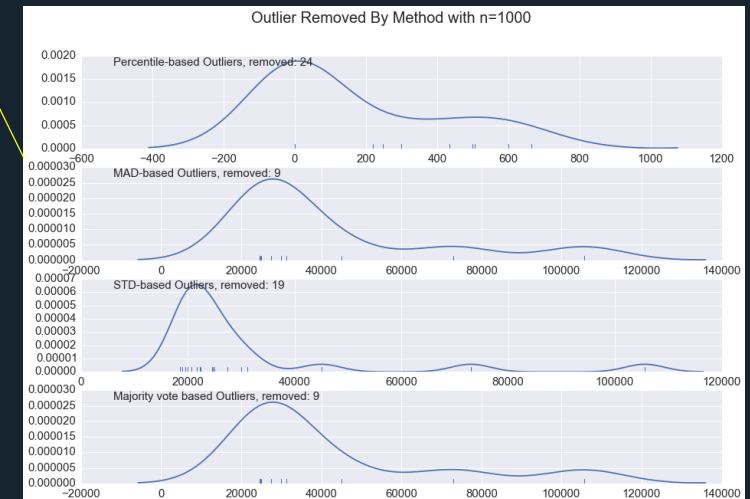
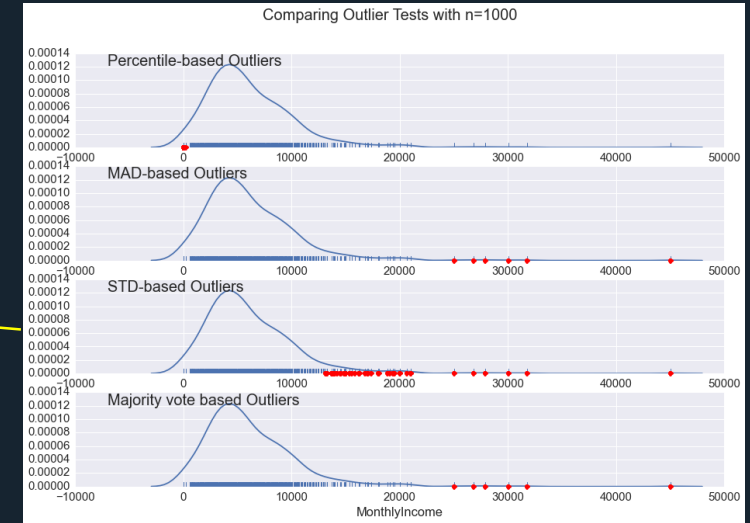
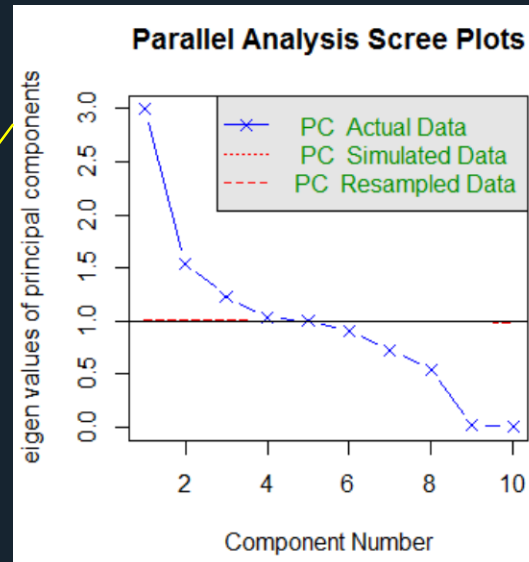


- "Monthly Income" has 29,731 NAs (~20%)
- "Number of Dependents" has 3,924 NAs (~3%)

Outliers

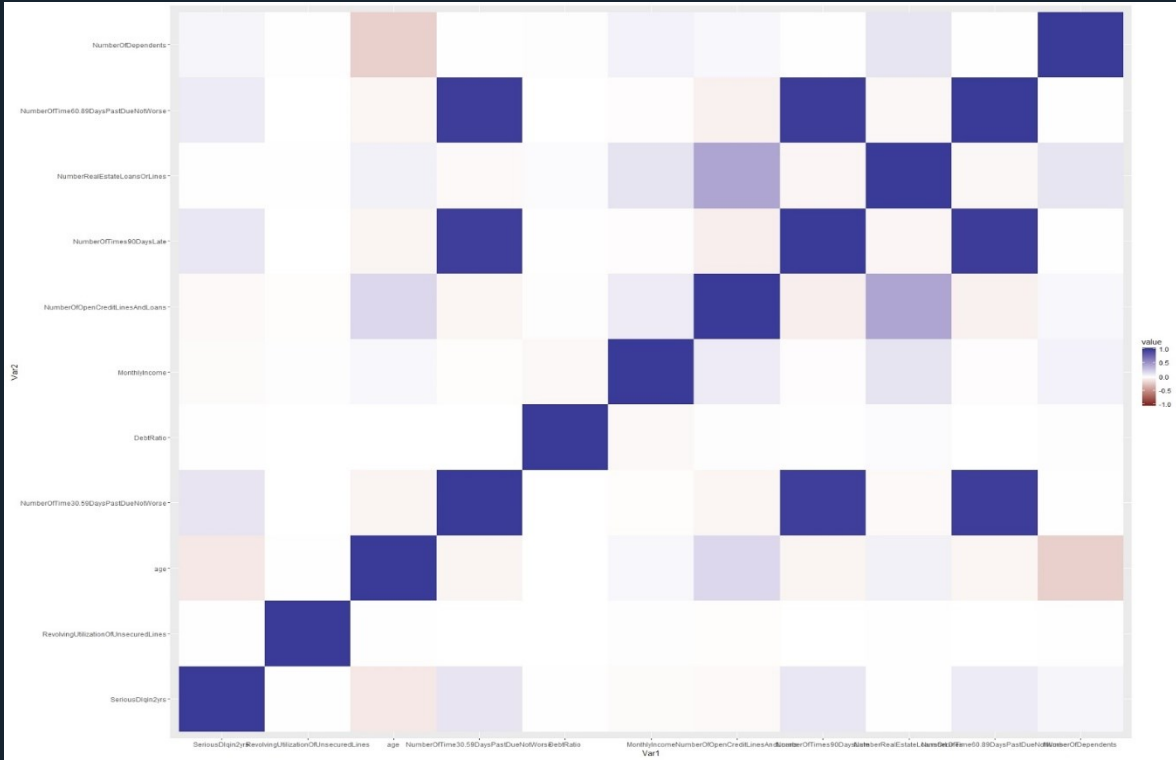
To filter or not to filter, that is a question... almost all variables have outlier problems

- PCA suggests choosing 5 principal components, which explains only 66% of the total variance
- We only have 10 features, thus PCA may not work very well

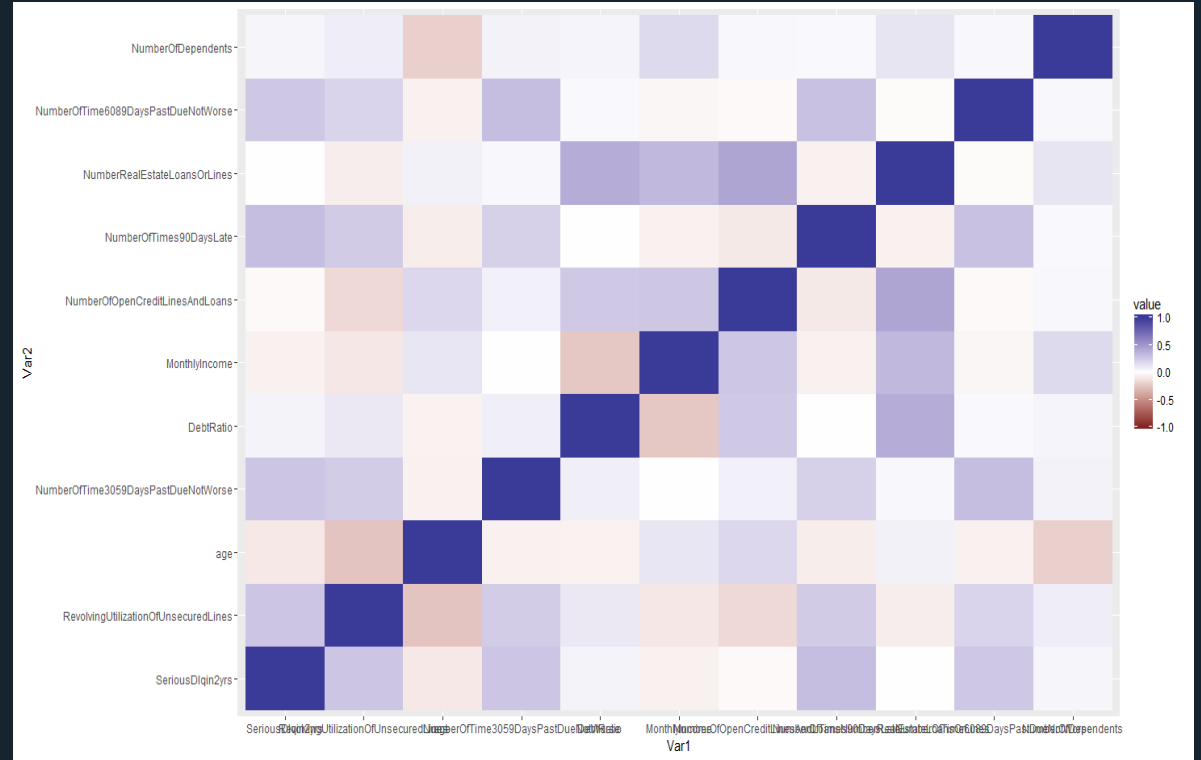


FEATURE CORRELATION HEATMAP

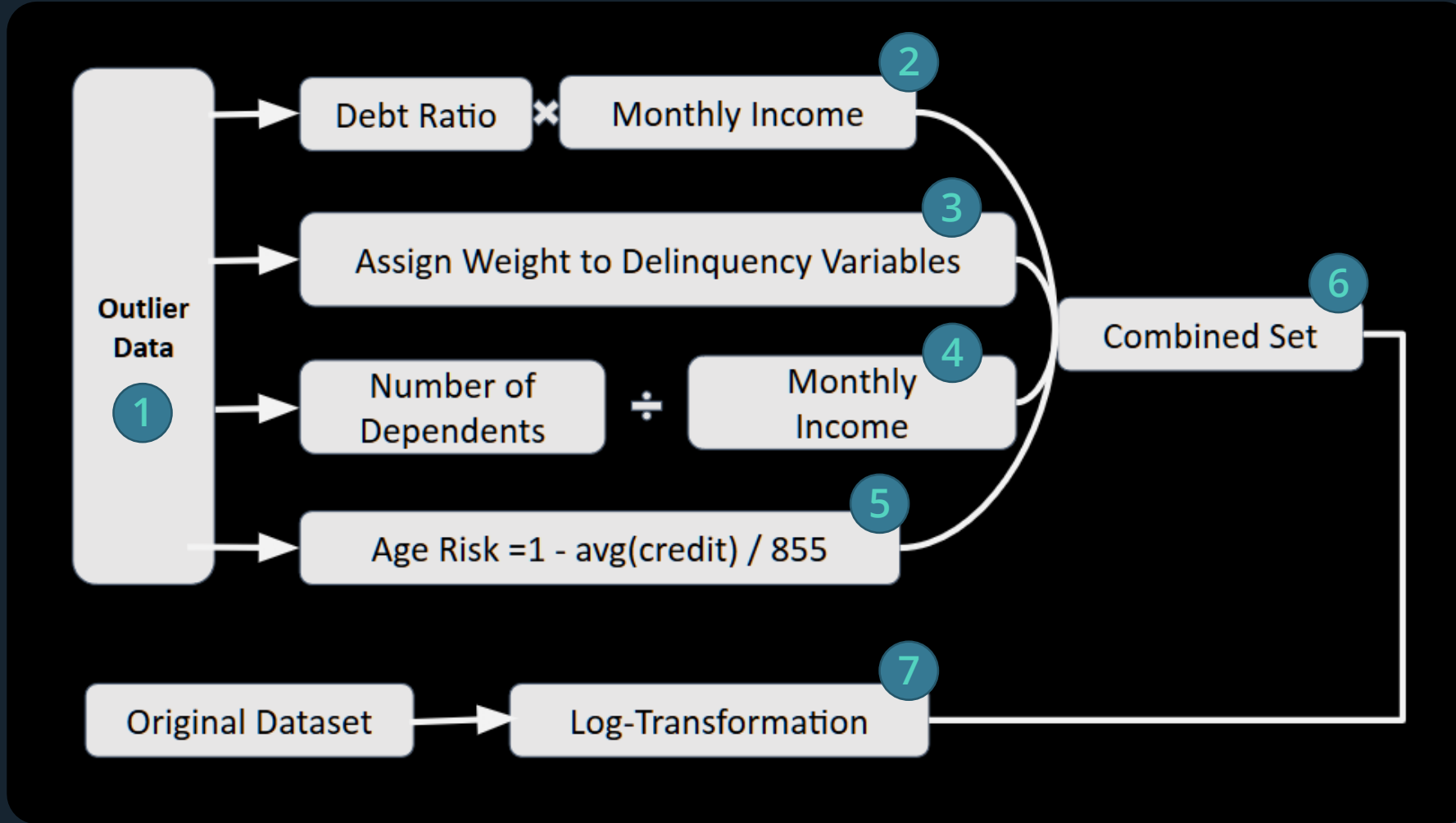
- Correlation no longer exists after the outliers are filtered

Number of 30-59
Days Past Due

Number of 90 Days Past Due

Number of 60-89
Days Past Due

FEATURE ENGINEERING



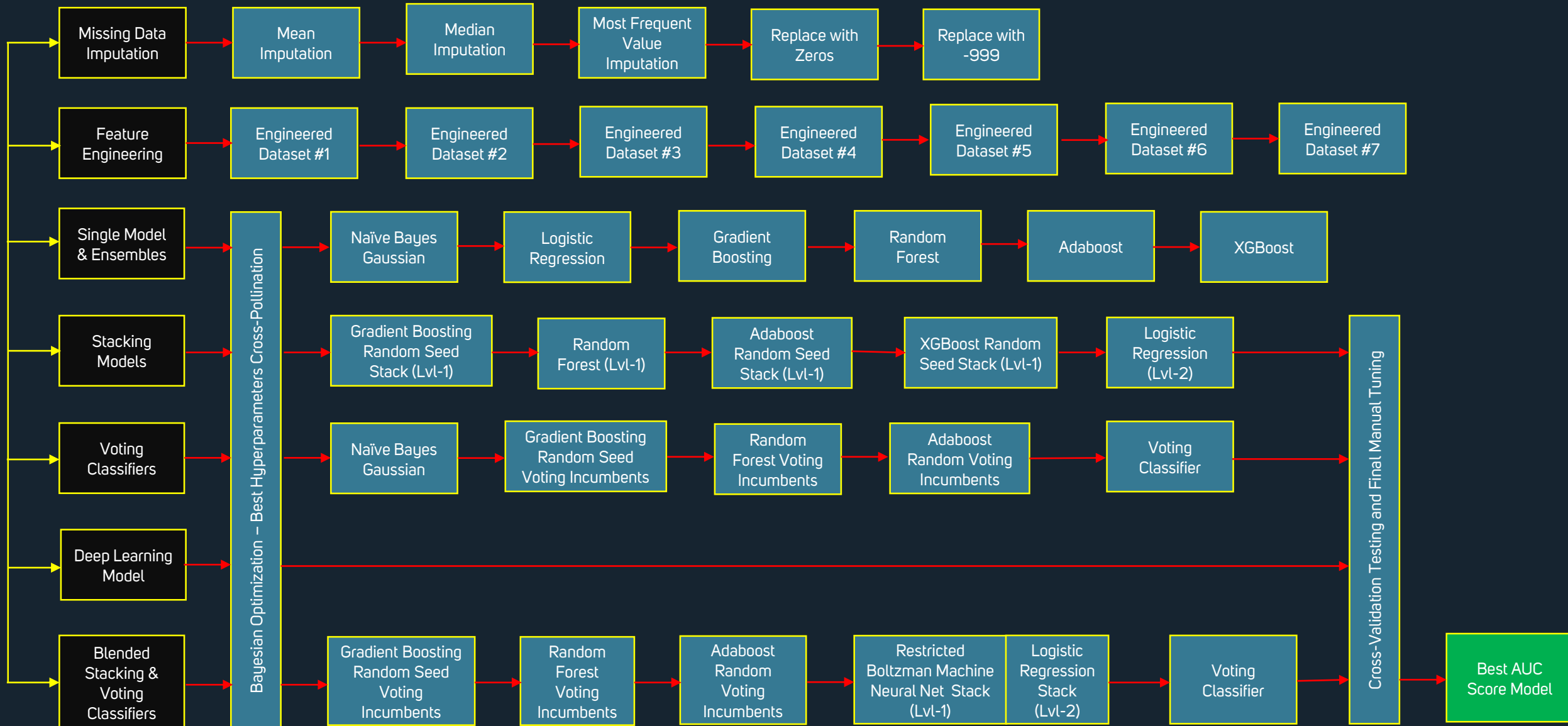
- Created 7 training data sets and 7 test datasets
- **Results:** Naive Bayes and Logistic regression AUC scores improved from ~0.7 to ~0.85
- However, for tree-based methods, the new features did not improve the scores

USE AGILE CHUNKING | FAIL FAST | KEEP ACCURATE LOGS | FOLLOW THE PROCESS

PATHWAY TO THE TOP SCORE



PARALLEL MODELING PROCESS





RESULTS & FINDINGS

- The models seemed to **perform well** without the missing values being imputed
- Stacking and Voting, and the **combination** of the two tend to have better predictive power compared to ensembles
- Feature Engineering improved the AUC score for single models (Naive Bayes and Logistic Regression) from ~0.7 to ~0.85 but did not have much impact on the Tree based methods
- The incremental increase in the predictive accuracy (AUC) is in the **order of 0.0001** as we move higher in Kaggle leaderboard - the tuning gets a lot harder



LESSONS & INSIGHTS

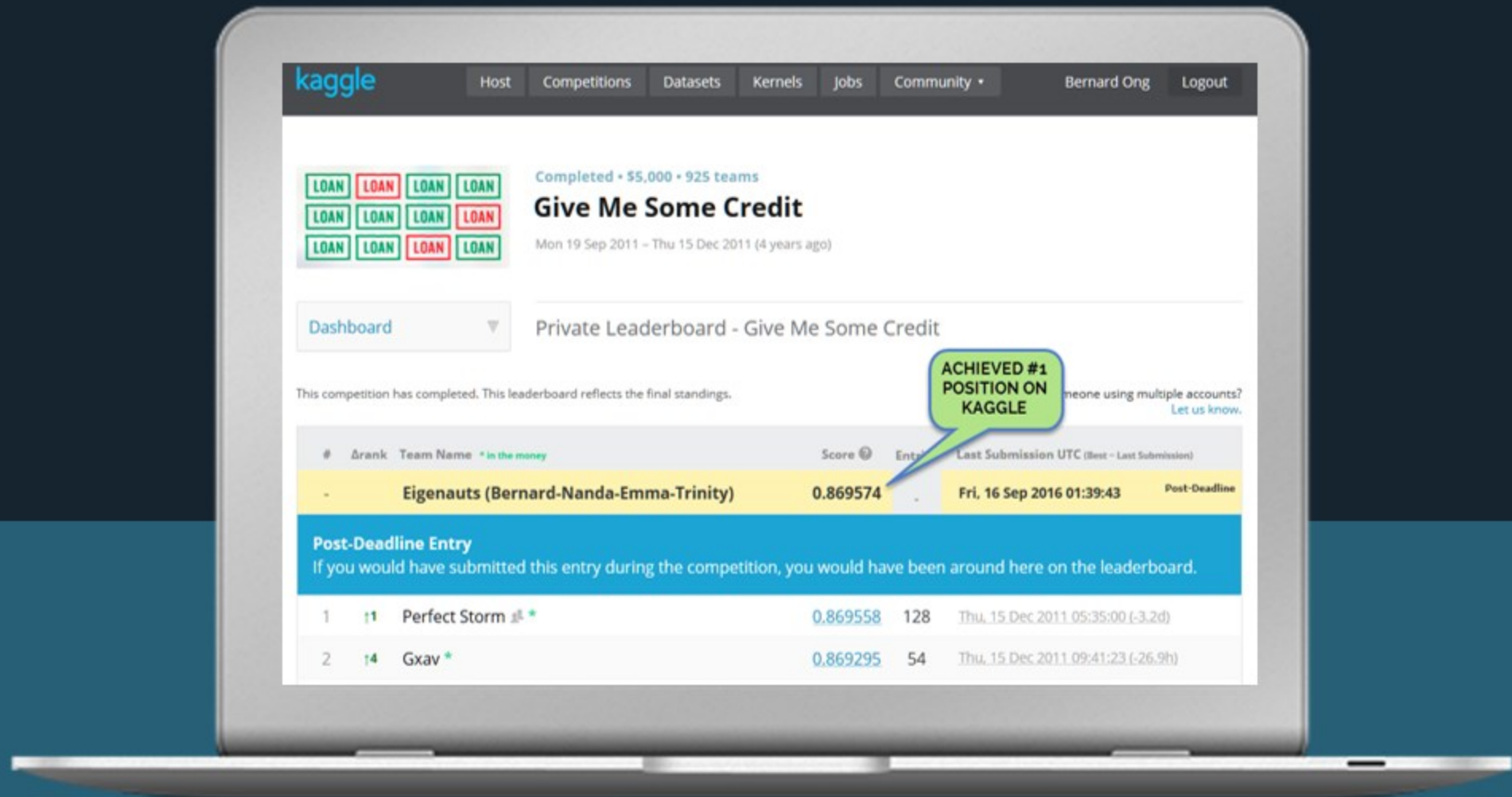
- Hyperparameter tuning is a very **time consuming process**
- Cross Validation is a **very critical** factor in determining model accuracy
- The model needs to be tuned at a much more **granular level** as the dataset gets smaller in size (i.e. number of features and observations)
- Following an **Agile process** has continued to be a proven factor for maximizing success



FUTURE STEPS

Explore how Theano / Keras
Deep Learning can be
integrated better into the Agile
modeling process.

Extend feature engineering to
add new polynomial and
transformed features.



TOP SCORE ACHIEVED

We not only surpassed our original goal to get on the Top 5% position, the team actually beat the high AUC ranking and achieved the #1 spot on the Kaggle challenge in 2 weeks.



MEET TEAM EIGENAUTS



BERNARD
ONG



EMMA (JIELEI)
ZHU

A PASSION FOR DATA

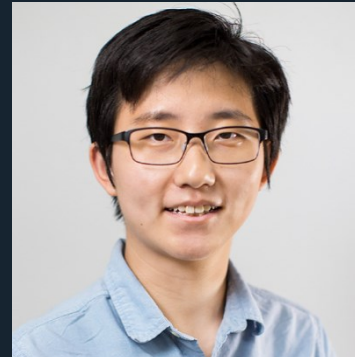
A dynamic team of Data Scientists that work very well together, ready to take on whatever challenges that come their way. The team is composed of an eclectic mix of a seasoned Executive, a Management Consultant, a creative problem solver and a theory-oriented problem solver. They do what it takes to excel in their chosen fields and are relentless in their pursuit of taking on innovative projects that exemplify their passion in working with data.



NANDA
RAJARATHINAM



TRINITY
(MIAOZHI) YU



NYC Data Science Academy – Summer 2016 Team