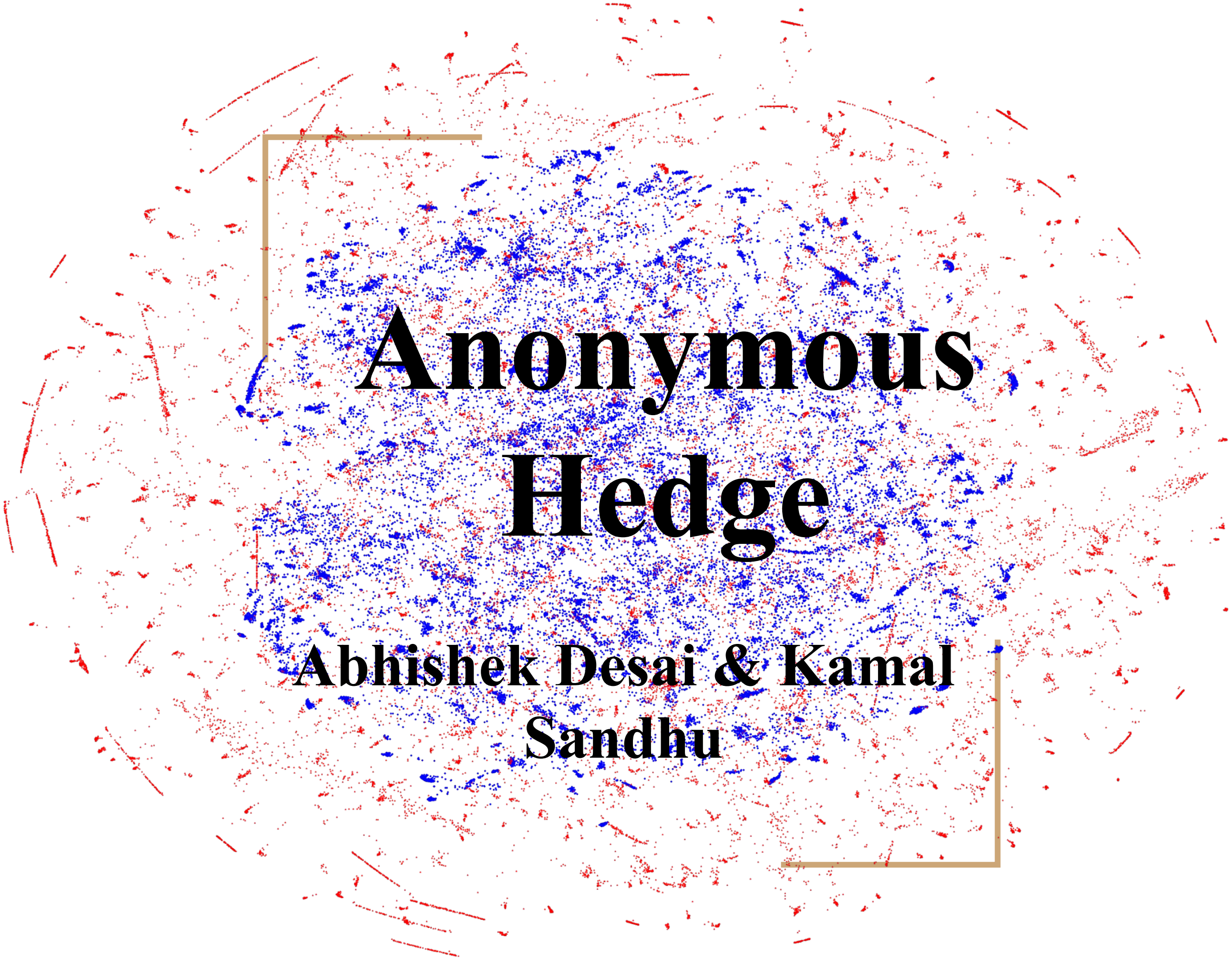
Anonymous Hedge



Abhishek Desai & Kamal Sandhu

Outline

Introduction

Concept

Methods applied

Results

Future

Numer.ai

- Hosts a weekly machine learning competition
- Predictions to be made on proprietary hedge fund financial data
- Somewhat like Kaggle but with a twist ..

Numer.ai

- One dataset every week
- Data is homomorphically encrypted
- Models trained on one week’s data are not relevant next week (encryption key changes from week to week)
- Presses contestants to design architectures that generalize from week to week

Homomorphic Encryption

- Encryption that allows computations to be carried out on ciphertext
- Generates an encrypted result which, when decrypted, matches the result of operations performed on plaintext

Homomorphic Encryption

- Numerai obtains predictions on its data without exposing the underlying data
- Machine learning predictions made on encrypted data are application to original data after reverse transformation

Homomorphic Encryption

Google homomorphic encryption if you want to know more

Why Encrypt Data?

Numerai argues -

- Financial markets are machine learning inefficient
- Small fraction of ML experts participate in it
- Open sourcing predictions will enable Numerai to benefit from this inefficiency
- Shares the profits with contestants

Downsides of Numerai

- Company is a black box
- Limited support provided for contestants
- Contestants cannot benefit from long term network effects or compounding
- Payout merely enough to make it worthwhile for contestants who can consistently place in top 10
- I still can’t figure out if it is a scam or not

Leaderboard Ranking

- 2 datasets - training (labeled) and tournament (used for scoring)
- Public score on a fraction of tournament dataset
- Numerai uses tournament data predictions submitted by contestants to build a “secret” meta model

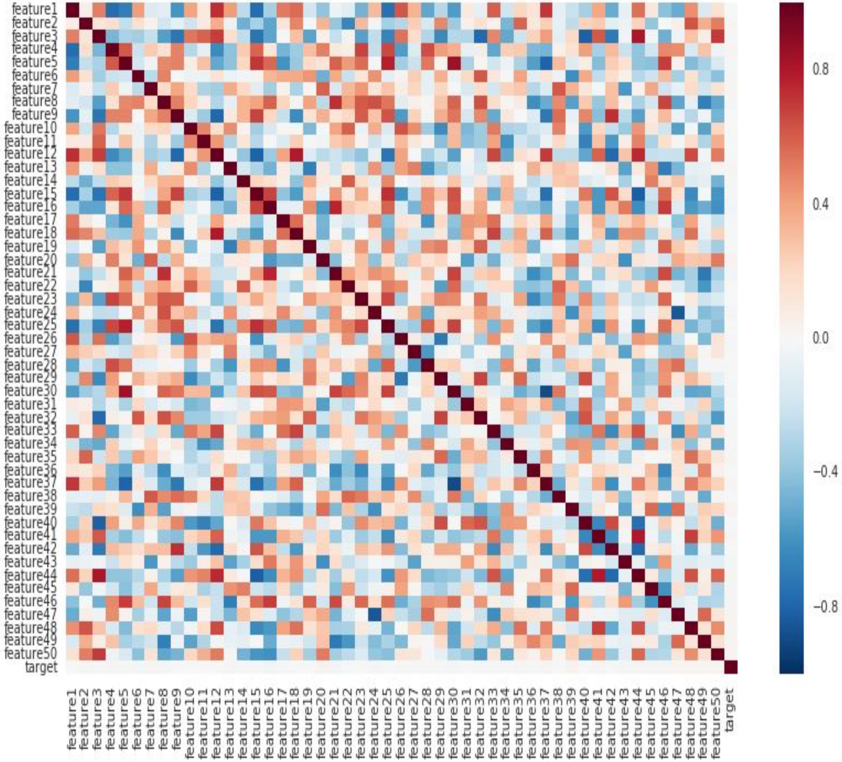
Leaderboard Ranking

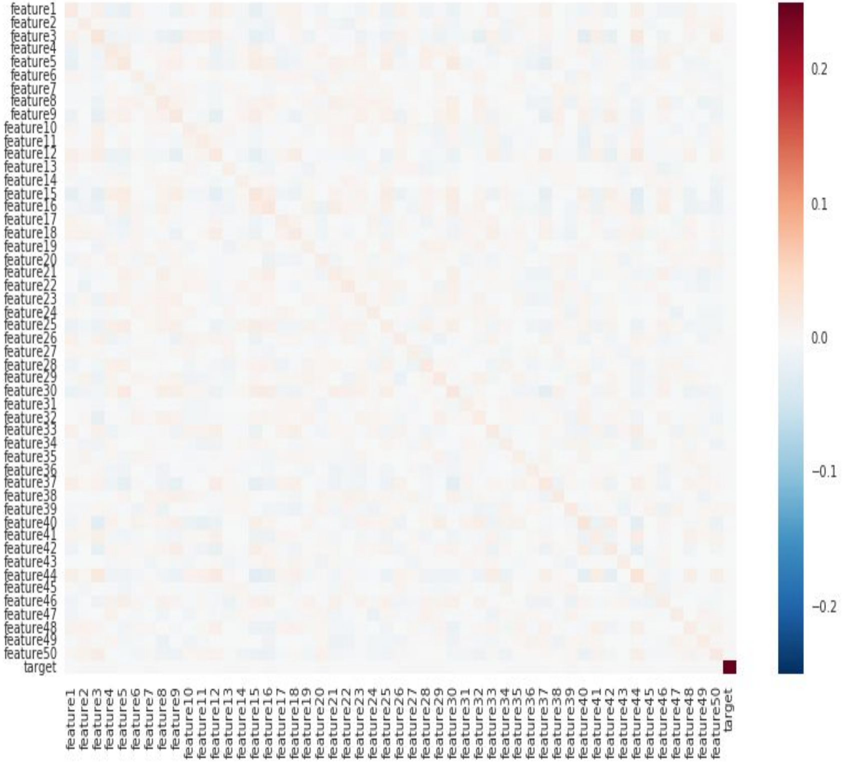
- Leaderboard ranking based on loss between “secret” meta model’s predictions and user submissions
- Can be thought of as a simpler game theory problem
- Making the right choice is not the best choice
- Best choice is to make the best choice after everyone else has made their right choices

Leaderboard Ranking

- In other words, best choice is to mimic Numerai’s meta model as it is the outcome of best individual submissions
- Kinda like the stock market but without price transparency

Correlation & Covariance Matrices





Basic Analysis & Ensemble

Used ‘Caret’ library in R

- Impute Missing Values(unnecessary)
- Split data 75%/25% for CV
- Define Training Controls
- Define Predictors and Outcome (feature for Classification, i.e ‘Target’)
- Build Models
    - Random Forest
    - KNN (21)
    - Logistic Regression
- Train & Predict Models individually using LogLoss as the evaluation metric

Basic Analysis & Ensemble - 3 Methods

Averaging: Since the predictions were classified (Y/N), averaging doesn’t make much sense for this binary classification. However, the observation probabilities were average-able as to whether they were going to be in either one of the binary classes

Majority Voting: In majority voting method, the assignment is made for the prediction based on the majority vote. Since there were 3 models for a binary classification, the issue of a tie was avoided. Random Forest won!!!

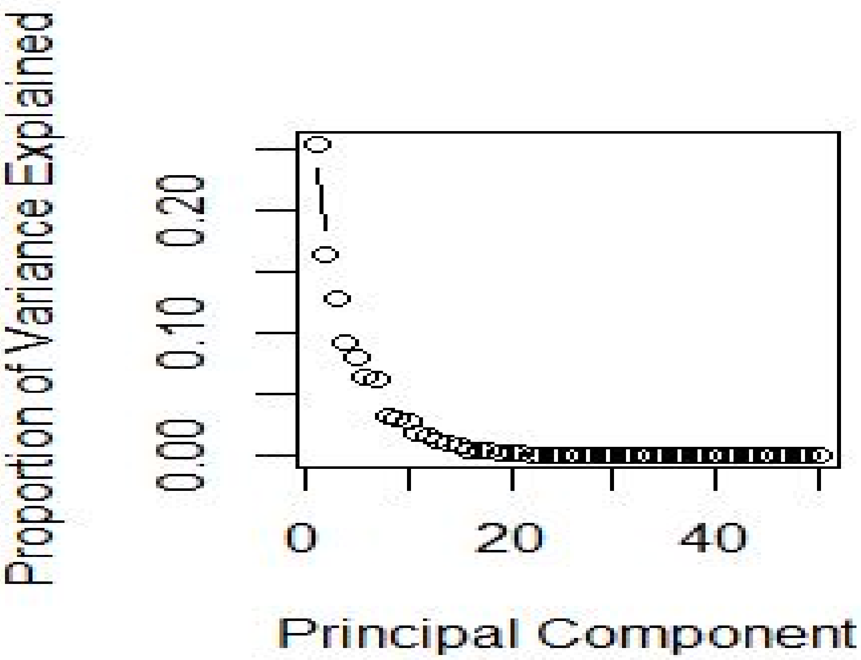
Weighted Average: As opposed to a simple average, a weighted average was also used with more accurate models carrying higher weights. 0.5 assigned to logistic regression and 0.25 to KNN and random forest each. Weighted Average logloss @ .694758

Basic Analysis & Ensemble - Stacking

We used a linear regression for making a linear formula for making the predictions in regression context for mapping base layer model predictions to the outcome or logistic regression for classification..

1. Train the individual base layer models on training data.
2. Predict using each base layer model for training data and test data.
3. Now train the top layer model again on the predictions of the bottom layer models that has been made on the training data.
4. Finally, predict using the top layer model with the predictions of bottom layer models that has been made for testing data.

NN (PCA & Raw Data) Stack Custom Algorithm



NN (PCA & Raw Data) Stack Algorithm

Build 2 Neural Networks with 2 hidden layers with 300 neurons in the first layer and 200 neurons in the 2nd layer

Use logloss as the cross-entropy loss function

Use softmax at the end since its a classification problem.

Feed the 1st NN with the Raw Data 49 neurons - logloss @ .6925

Feed the 2nd NN with the Principal Components (21 neurons, based on Scree Plot)

Stack them together

Use a lasso on the combined data set to devise a model with some condensed data and some normal data

Best Model

1. K-means with 2, 4, 16, 64 and 128 clusters
2. 8 sklearn models
   a. Combination of linear, decision trees, XGB, NN and ensemble models
   b. Tuning using Bayesian hyperparameter optimization
3. Deep neural net with 4 hidden layers
   a. SGD optimizer
   b. Tanh activation and uniform initialization
   c. 0.25 dropout and batch normalization after every layer

Best Model

4. Ensembled using a soft voting criteria

5. Eliminated individual models based on -

    a. Time and cost complexity
    b. That gave a log-loss of more than 0.6931 (worse than guessing)

Ensembling

Other ensembling methods tried -

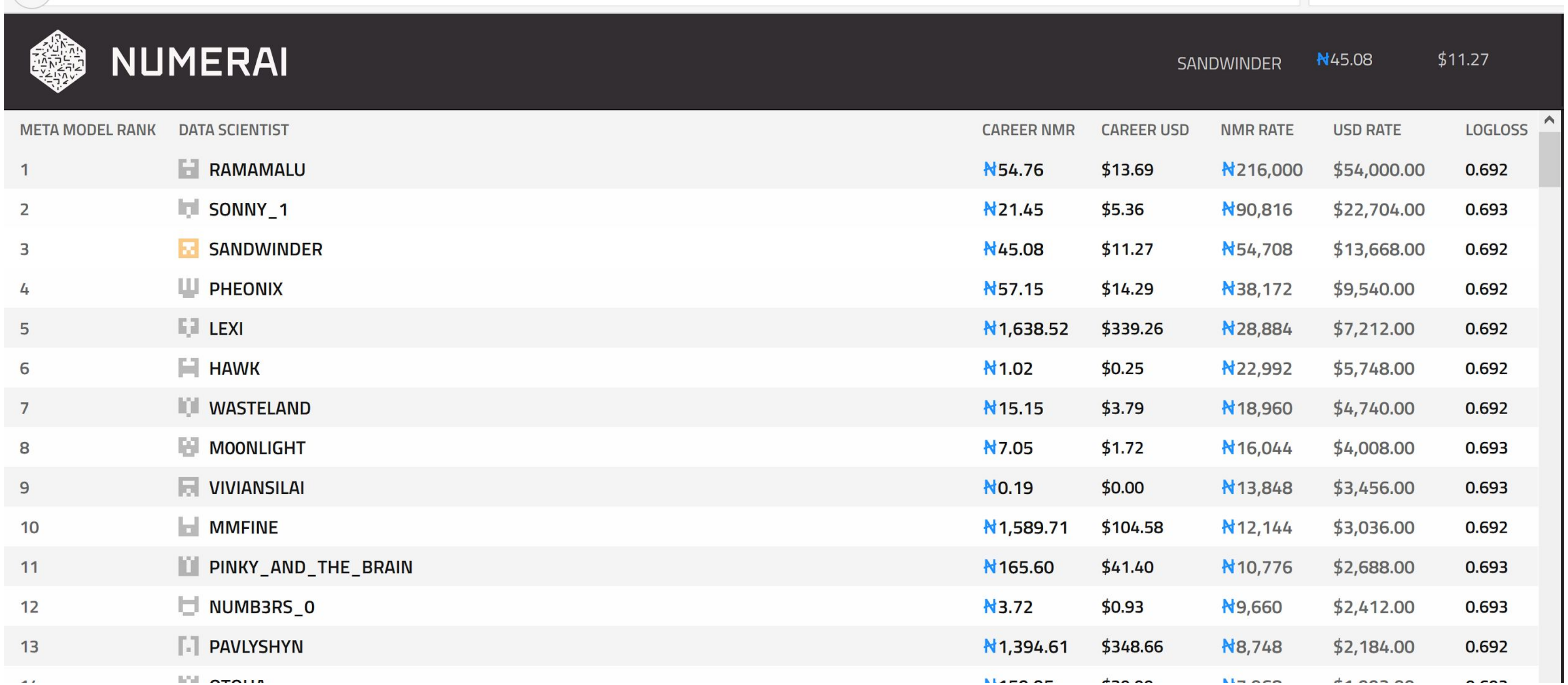
- Stacking with features
  - By adding predictions from upstream models as additional features
  - 4 hidden layer neural net as a meta classifier
- Stacking without features
  - Predictions based on predictions of upstream models
  - 1 hidden layer neural network as a meta classifier
- Hard voting classifier
  - Averaging predictions of all the models

Ensembling

Models that sunk to the bottom -

- Soft Voting classifier with probabilities calibrated using the Calibrated Classifier CV from sklearn
- Multi-level perceptron in Keras
- Both had the same log-loss as the other models

Results



NUMERAI

SANDWINDER ₦45.08 $11.27

META MODEL RANK	DATA SCIENTIST	CAREER NMR	CAREER USD	NMR RATE	USD RATE	LOGLOSS
1	RAMAMALU	₦54.76	$13.69	₦216,000	$54,000.00	0.692
2	SONNY_1	₦21.45	$5.36	₦90,816	$22,704.00	0.693
3	SANDWINDER	₦45.08	$11.27	₦54,708	$13,668.00	0.692
4	PHEONIX	₦57.15	$14.29	₦38,172	$9,540.00	0.692
5	LEXI	₦1,638.52	$339.26	₦28,884	$7,212.00	0.692
6	HAWK	₦1.02	$0.25	₦22,992	$5,748.00	0.692
7	WASTELAND	₦15.15	$3.79	₦18,960	$4,740.00	0.692
8	MOONLIGHT	₦7.05	$1.72	₦16,044	$4,008.00	0.693
9	VIVIANSILAI	₦0.19	$0.00	₦13,848	$3,456.00	0.693
10	MMFINE	₦1,589.71	$104.58	₦12,144	$3,036.00	0.692
11	PINKY_AND_THE_BRAIN	₦165.60	$41.40	₦10,776	$2,688.00	0.693
12	NUMB3RS_0	₦3.72	$0.93	₦9,660	$2,412.00	0.693
13	PAVLYSHYN	₦1,394.61	$348.66	₦8,748	$2,184.00	0.692

Results

- Stayed in top 10 most of the time
- Although rank fluctuated between 3rd and 25th
- Models submitted earlier in the week tend to slide down over time
- Made our first $$$ from data science

Future/Learning

- Fill the gaps. Don’t leave anything on the table
- Automation of machine learning using -
  - Notebooks and maybe a package (contribution or new)
  - Preprocessing steps
  - Algorithm selection
  - Bayesian hyperparameter optimization
  - Benchmarking models
- Transfer learning from past dataset