# Kaggle Competition:
# Allstate Claims Severity

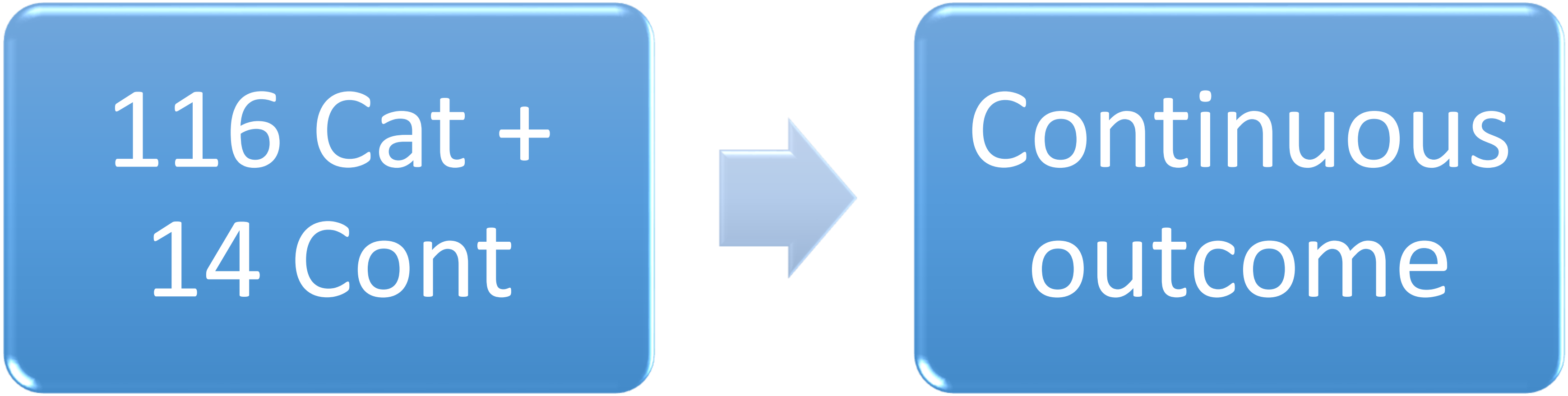
**Group XBX:**
**Andrew, Alex, David**

kaggle™

Allstate®
You're in good hands.

# 1 Introduction

Machine Learning Competition



116 Cat + 14 Cont

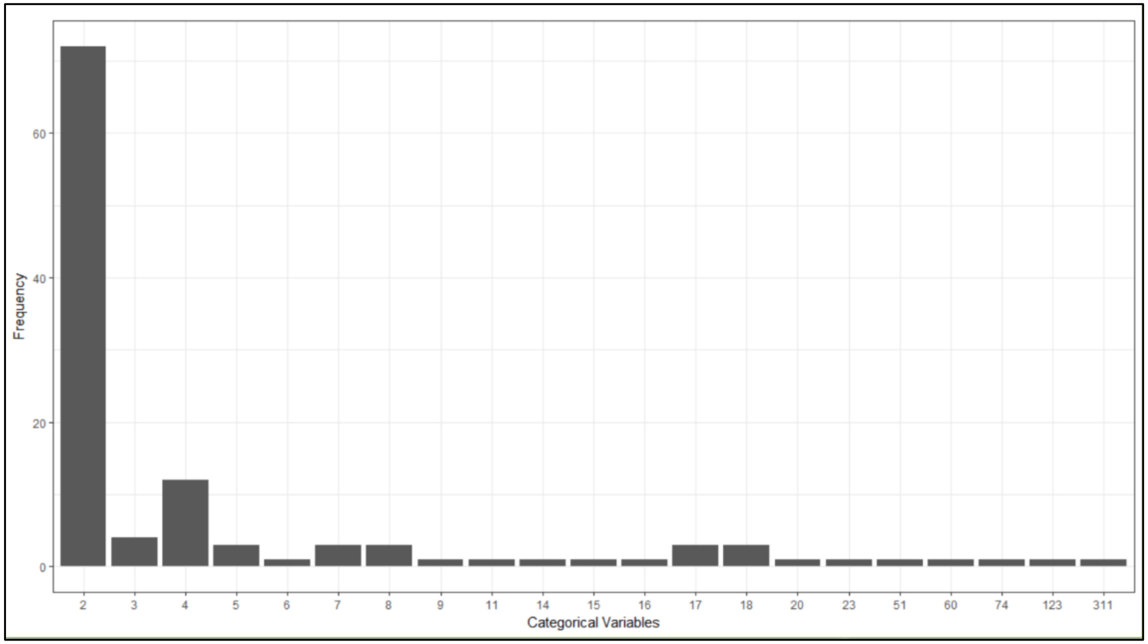
Continuous outcome

Goal: Create algorithms to minimize MAE

# 2 Variable EDA

- Log Transformation on the prediction variable *loss*
- Level of categorical variables ranges from 2 to 311, crucial for feature selection

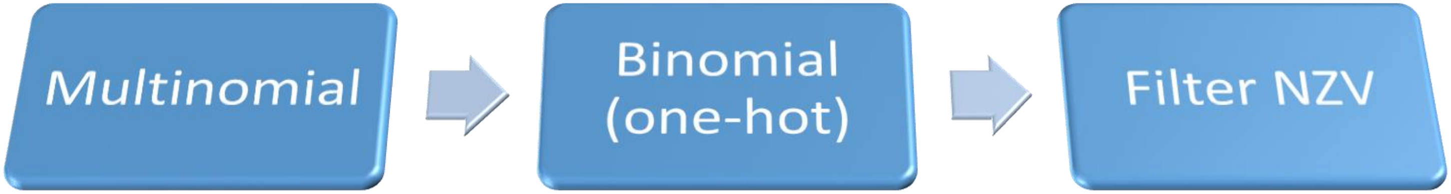
| Level | # of Categorical Vars |
|---|---|
| 2 | 72 |
| 3 | 4 |
| 4 | 12 |
| … | .. |
| 74 | 1 |
| 123 | 1 |
| 311 | 1 |



- Possible sigmoid transformation on continuous variables

# 3.1 Encoding 1

Idea:

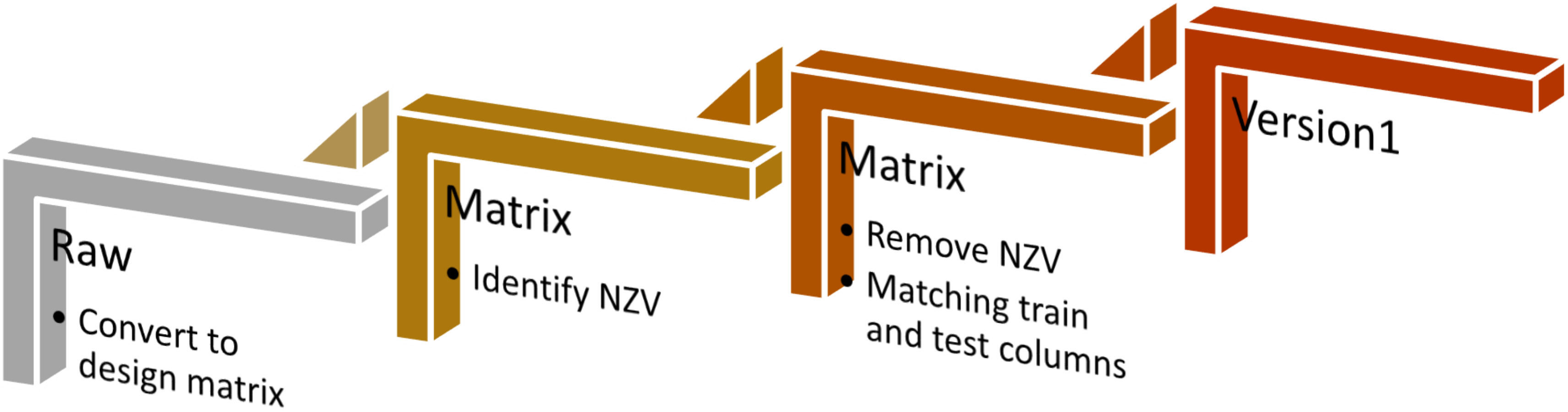


Multinomial

Binomial (one-hot)

Filter NZV

Process (caret package):



Raw

- Convert to design matrix

Matrix

- Identify NZV

Matrix

- Remove NZV
- Matching train and test columns

Version1

After: 193 variables

# 3.2 Encoding 2

Idea:

Combine → Re-level → To Numeric

Process (base R):

Combine

- *1. Convert to character*
- *2. Row bind train and test*

Compare

- Find different levels between train and test

Re-level

- Set all unmatched levels to NA
- Re-factorize

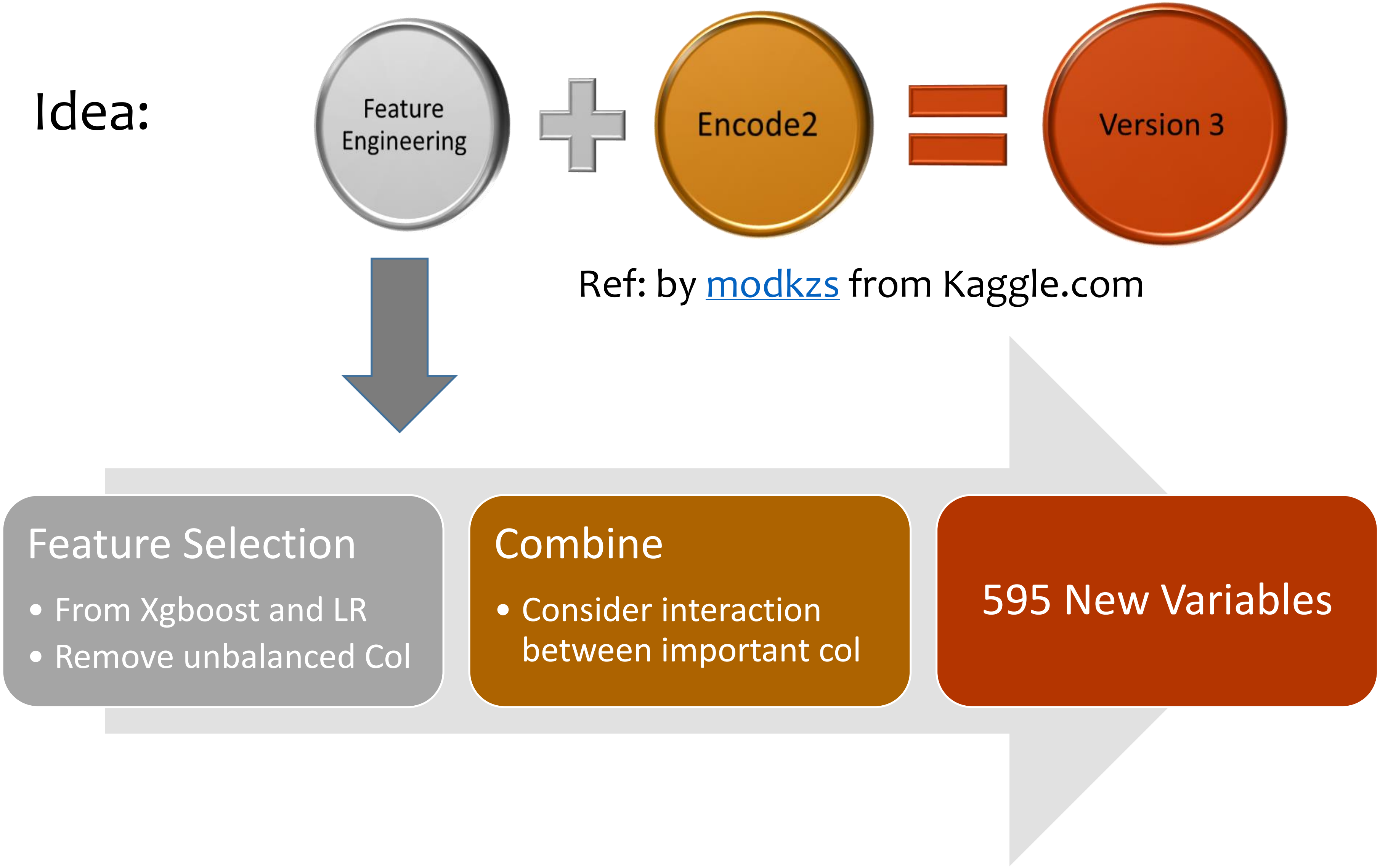
Version 2

- Re-Split

After: 130 variables

# 3.3 Encoding 3

Idea:



Feature Engineering

Encode2

Version 3

Ref: by modkzs from Kaggle.com

Feature Selection

- From Xgboost and LR
- Remove unbalanced Col

Combine

- Consider interaction between important col

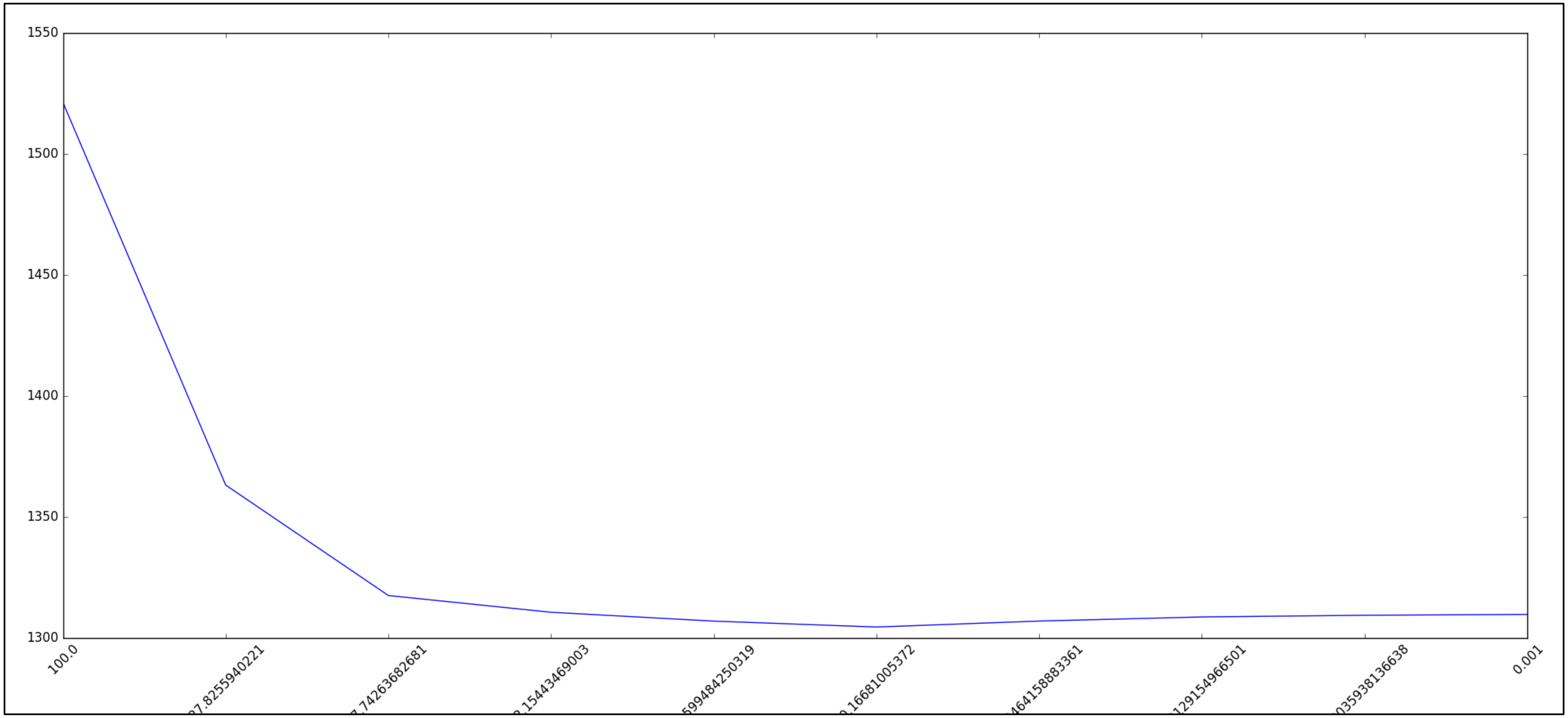
595 New Variables

After: 725 variables

# 4 Exploratory Model Selection

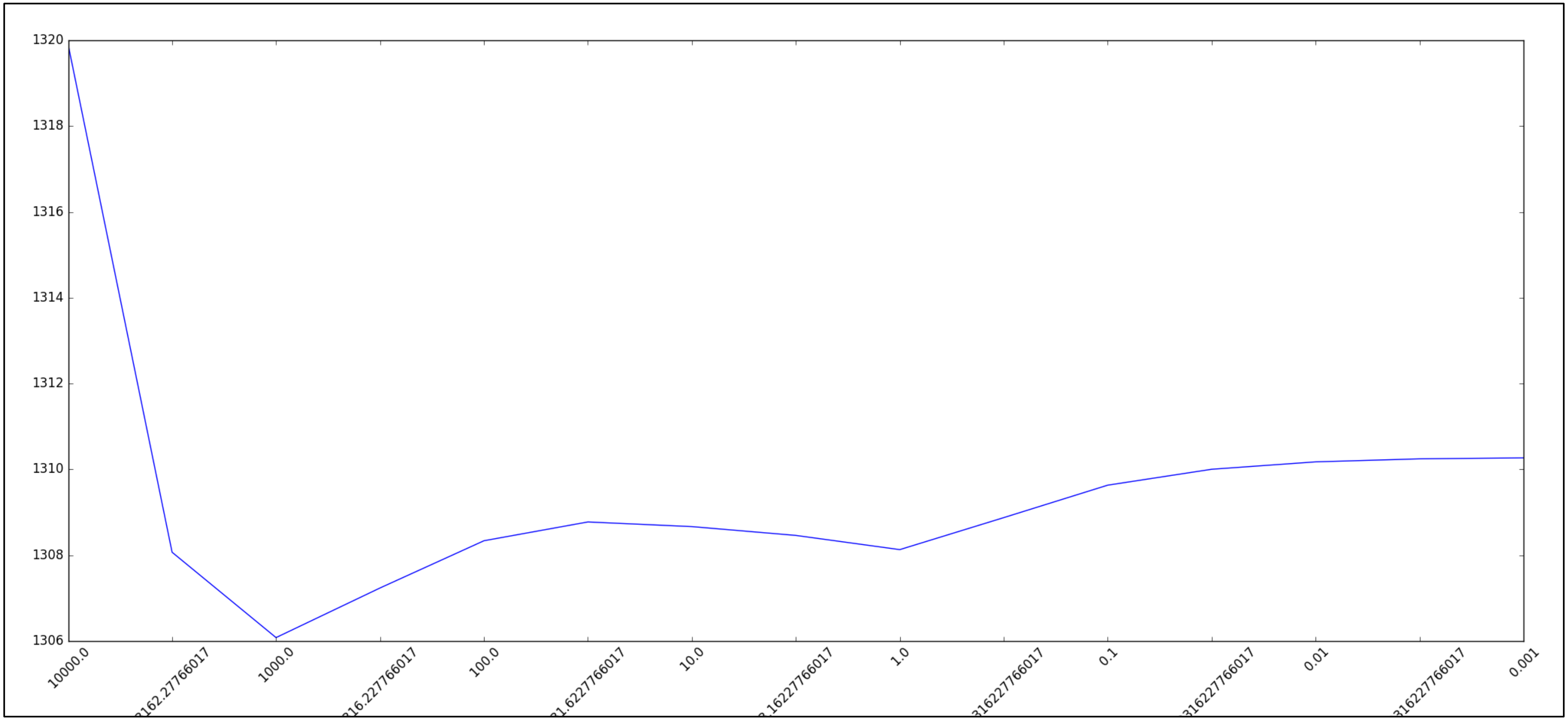
- Encoding Method: One Hot Encoding Unfiltered
- Validation: Train-Test Split (90-10)
- Machine Learning Algorithms:
    - LASSO Regression
    - Ridge Regression
    - CART
    - Random Forest
    - XGBoost
    - Neural Network

# 4.1 LASSO Regression



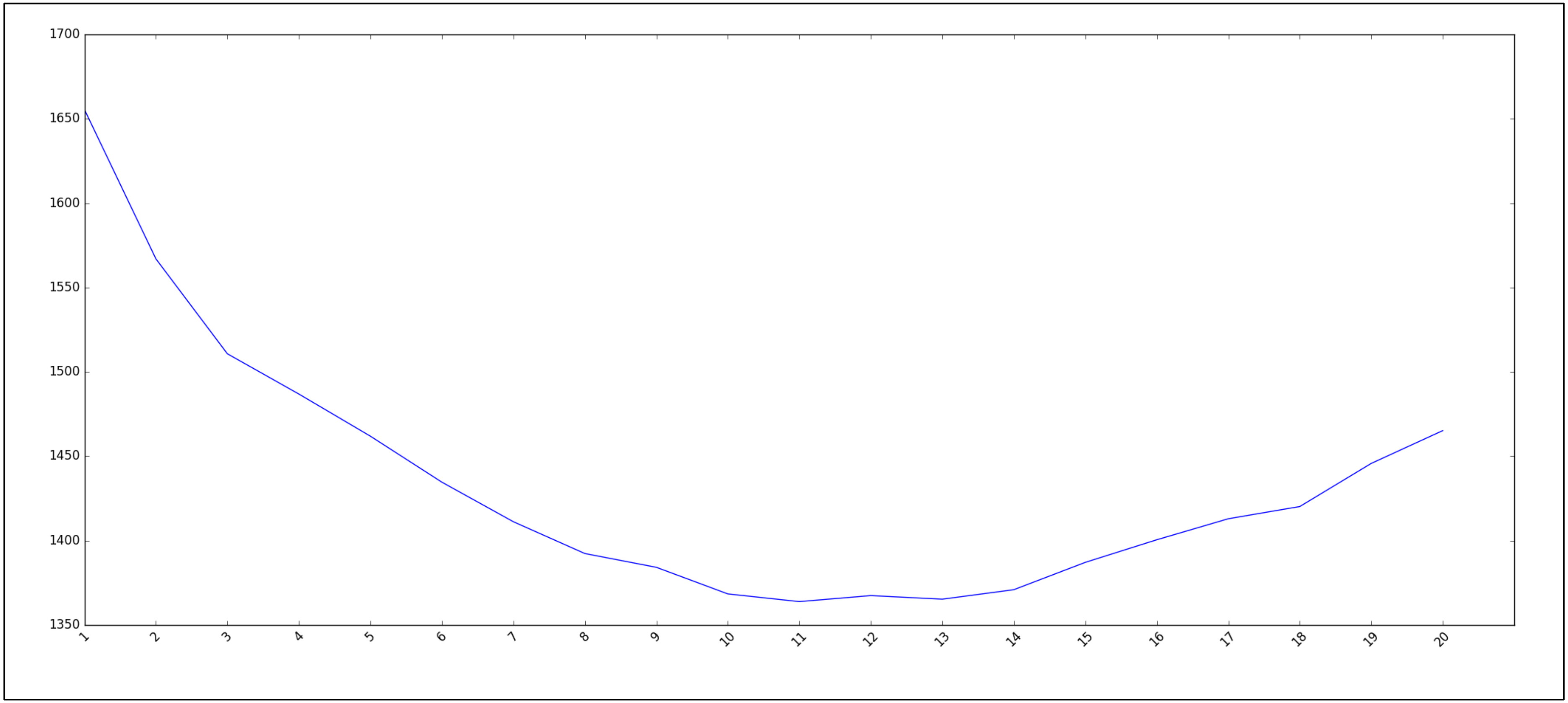
- Best lambda: 0.16681005372
- Mean Absolute Error: 1304.562

# 4.2 Ridge Regression



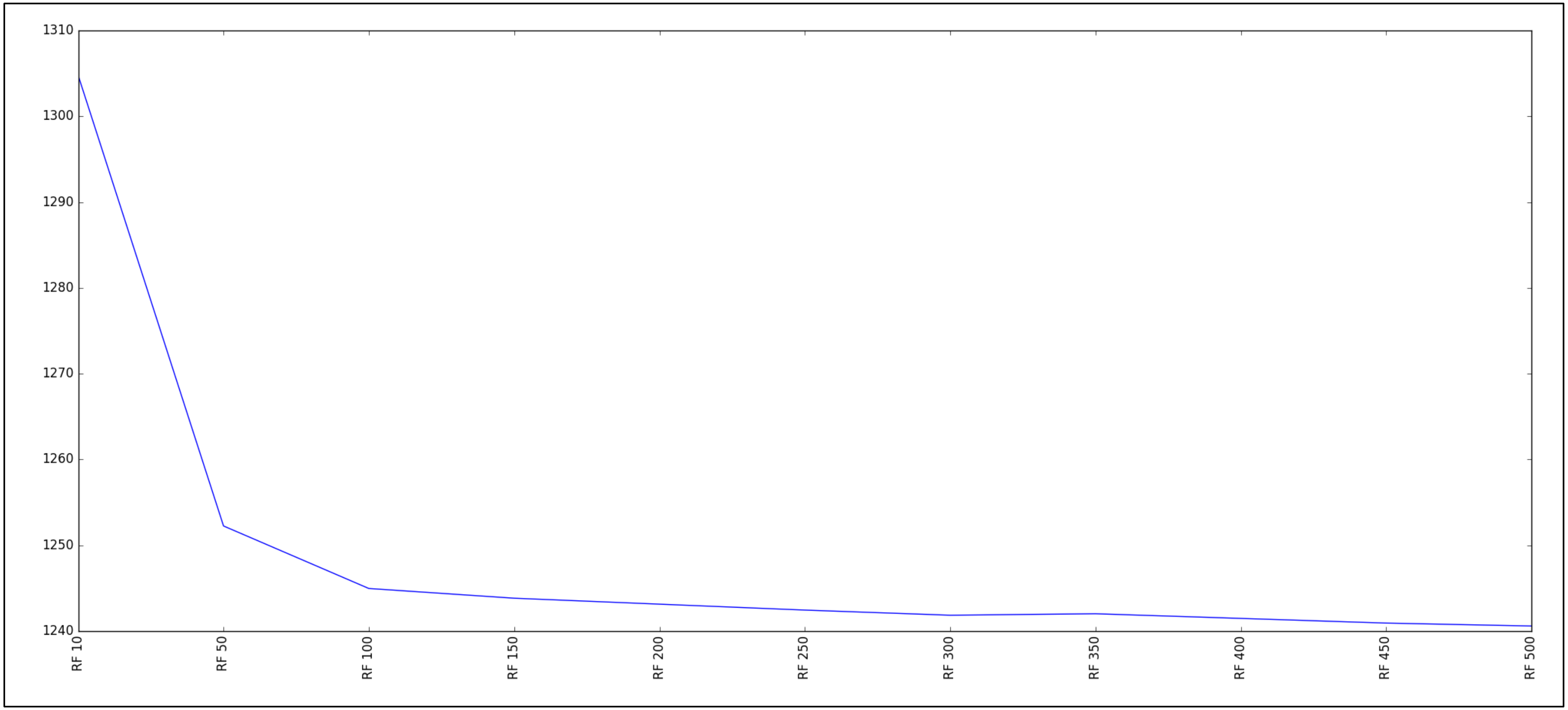
- Best lambda: 1000.0
- Mean Absolute Error: 1306.081

# 4.3 CART



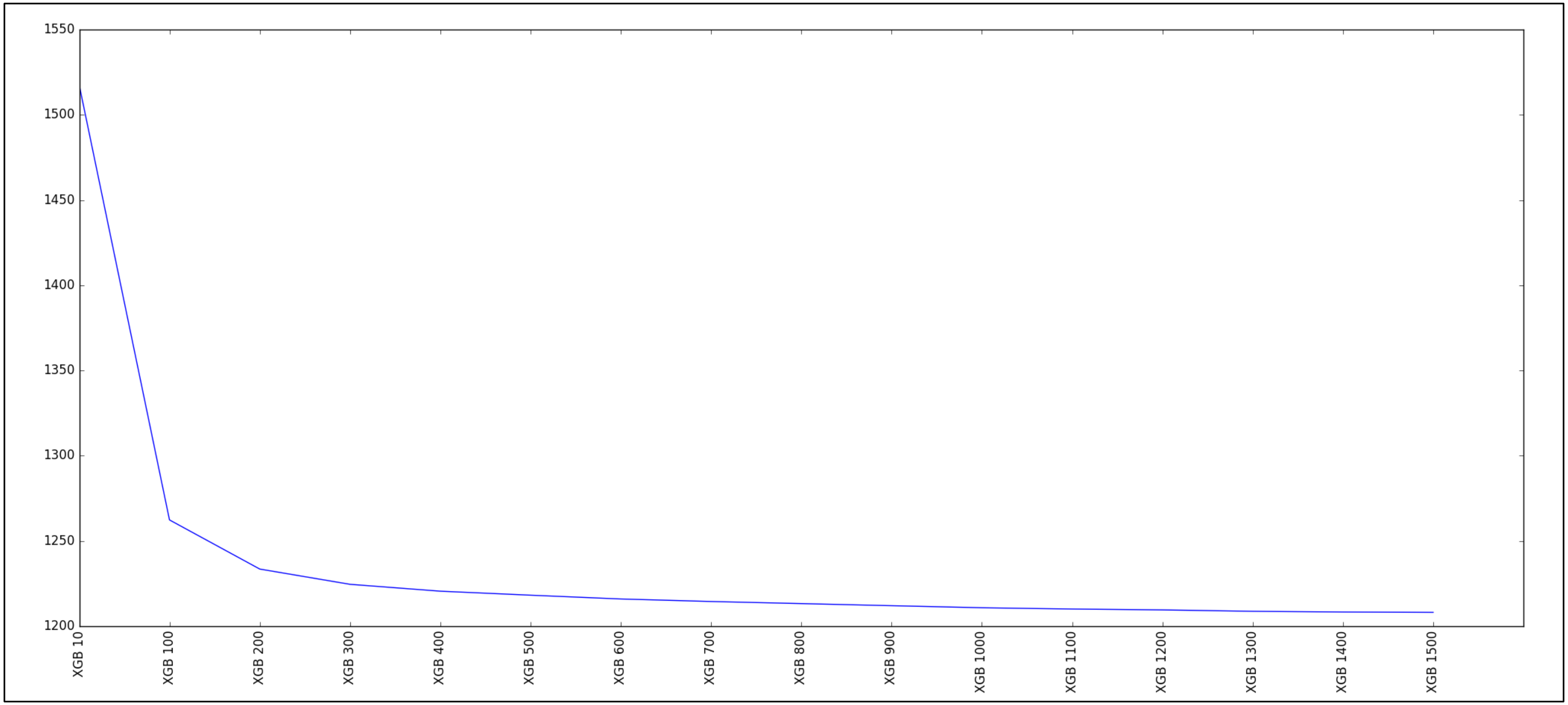
- Best Tree Depth: 11
- Mean Absolute Error: 1363.859

# 4.4 Random Forest



- Number of Trees: 500
- Mean Absolute Error: Approaching 1240

# 4.5 XGBoost



- Number of boosted models: 1500
- Mean Absolute Error: Approaching 1200

# 4.6 Exploratory Model Selection Summary

| Model | Mean Absolute Error |
|---|---|
| LASSO | 1304.562 |
| Ridge | 1306.081 |
| CART | 1363.858 |
| Random Forest | ~1240 |
| XGBoost | ~1200 |
| Neural Network | - |

- Best Models: Random Forest, XGBoost, Neural Network

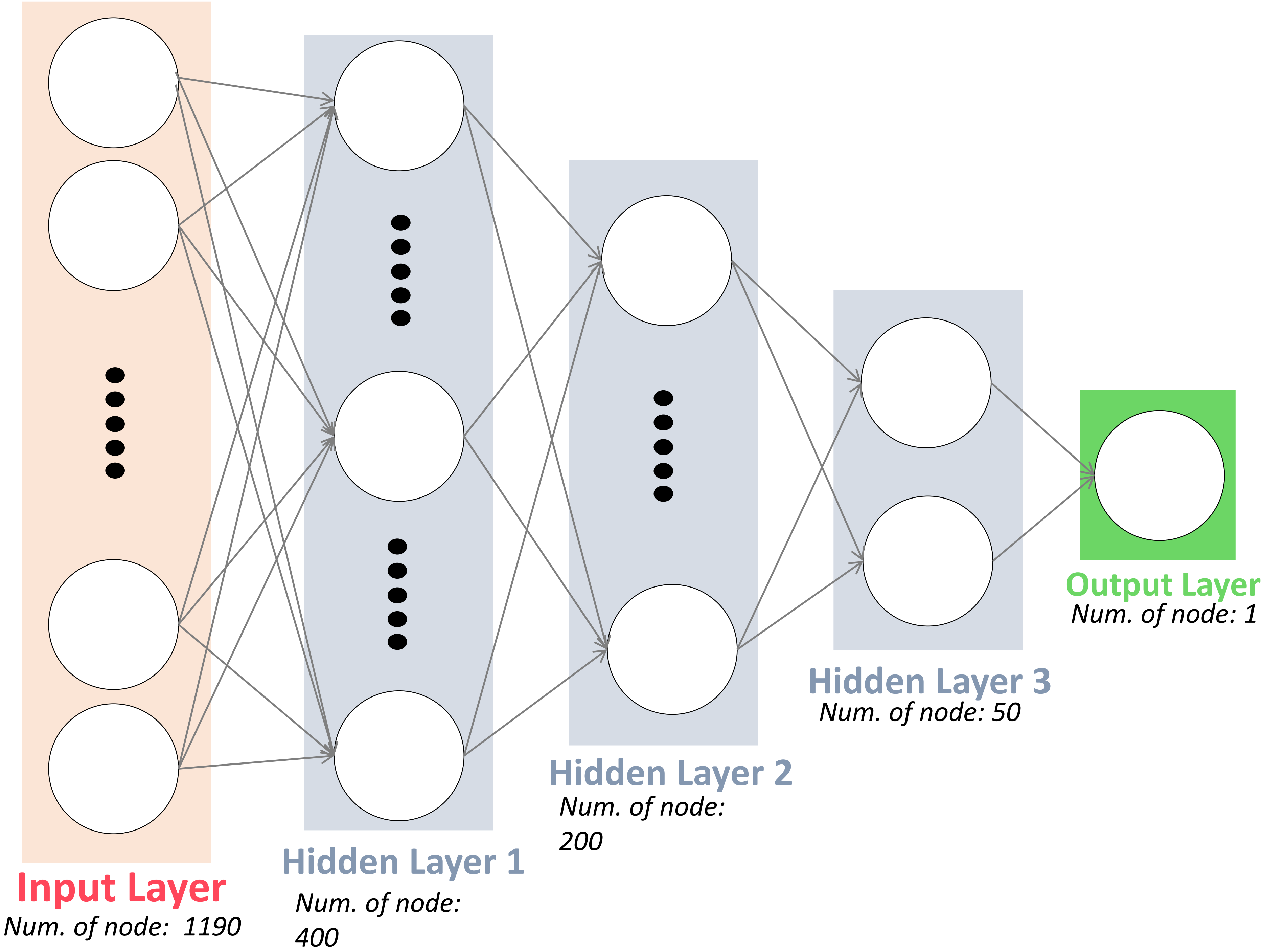
# 4.7 Neural Network

## Work Flow

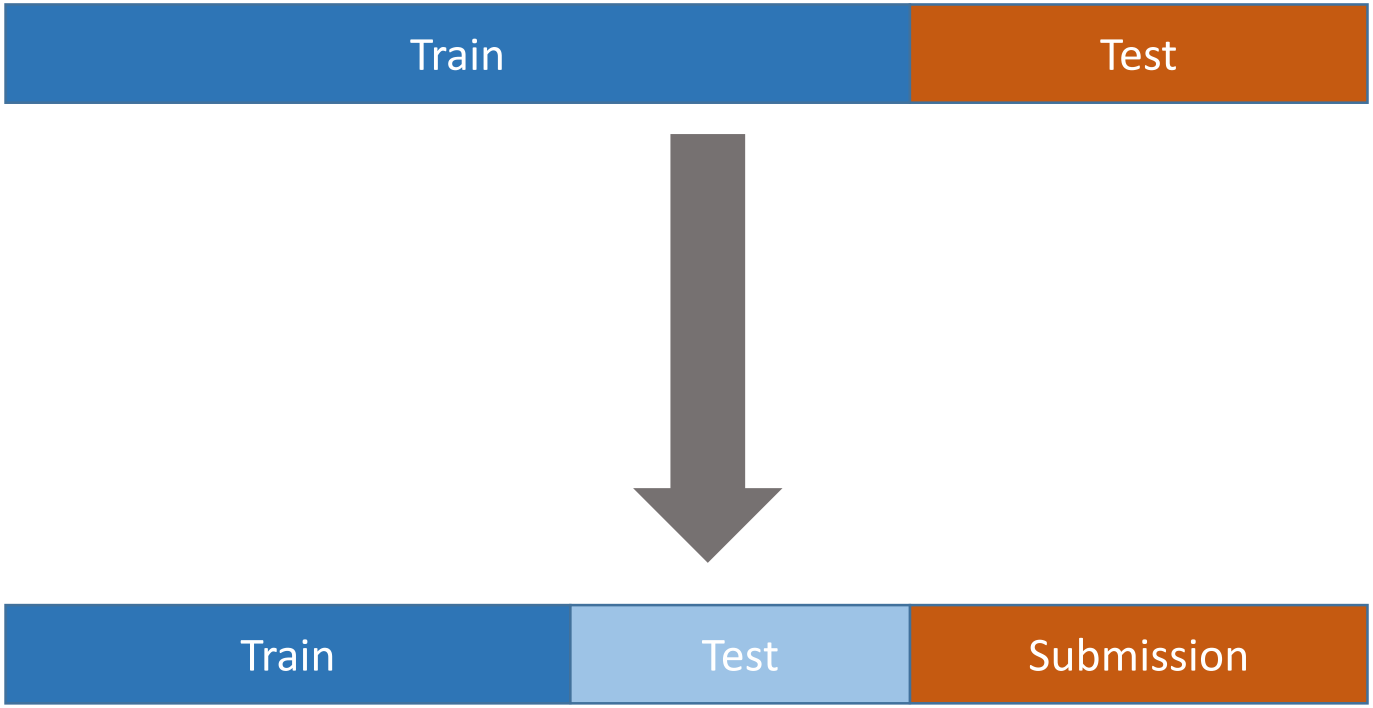
1. Create dummy variables
2. Put all dummitized variables into sparse matrix
3. Scale numerical variables for normalization
4. Setting parameters of the neural net construction
5. Stack using 10-fold cross-validation and bagging.
6. Use mean absolute error to evaluate.

## Key Parameters

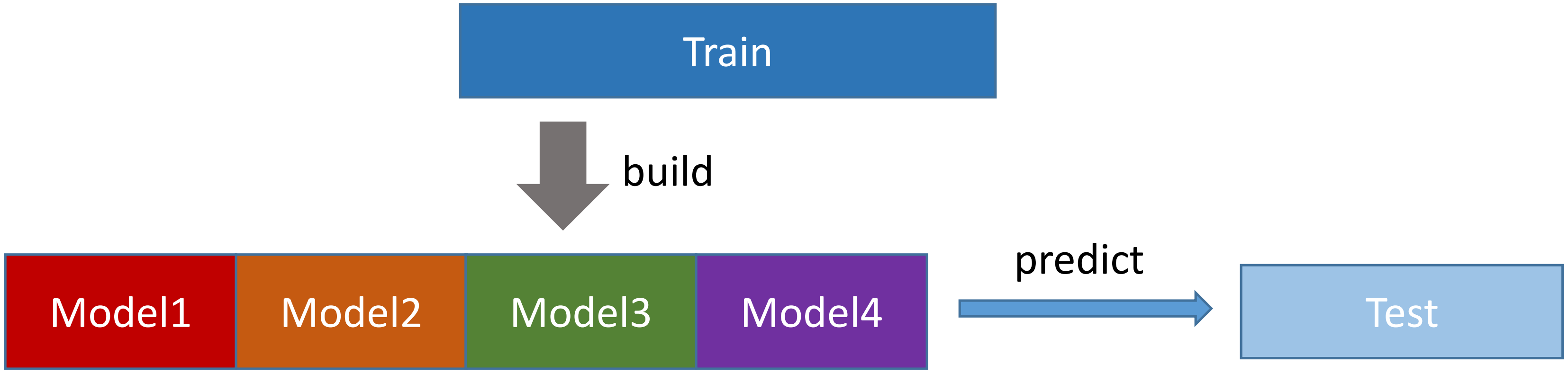
1. Number of bags
   - Number of bags used for bagging
2. Number of epochs
   - Number of times all of the training vectors are used once to update the weights
3. Dropout


Input Layer
Num. of node: 1190
Hidden Layer 1
Num. of node: 400
Hidden Layer 2
Num. of node: 200
Hidden Layer 3
Num. of node: 50
Output Layer
Num. of node: 1

# 5.1 Workflow for Stacking



# 5.1 Workflow for Stacking



Train

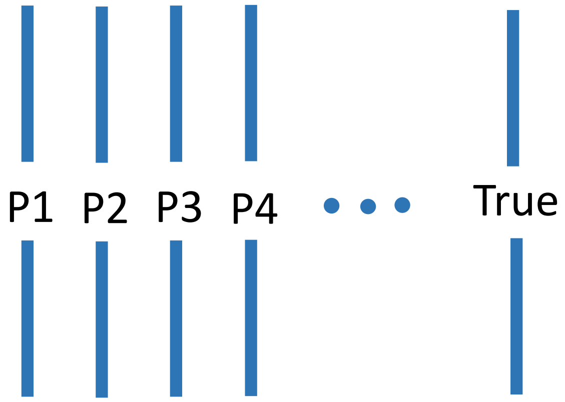
build

Model1 Model2 Model3 Model4

predict

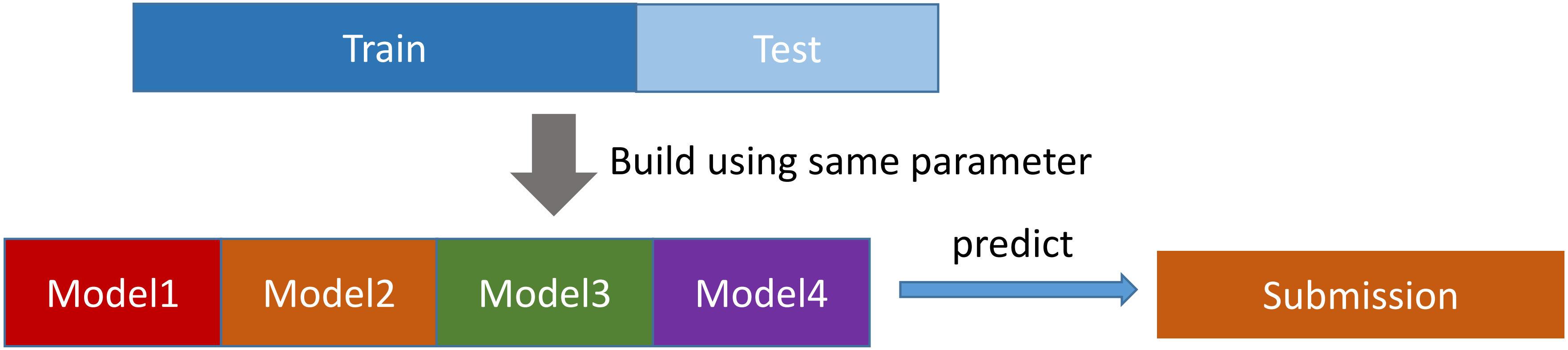
Test

We get:



P1 P2 P3 P4 • • • True

# 5.1 Workflow for Stacking



Train

Test

Build using same parameter

Model1

Model2

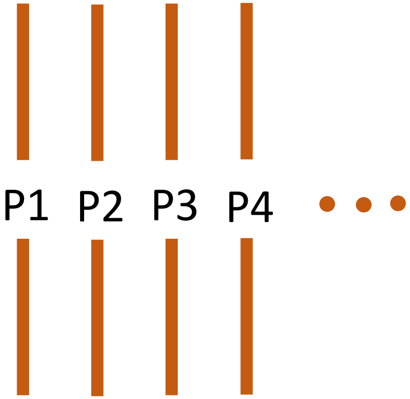
Model3

Model4

predict

Submission

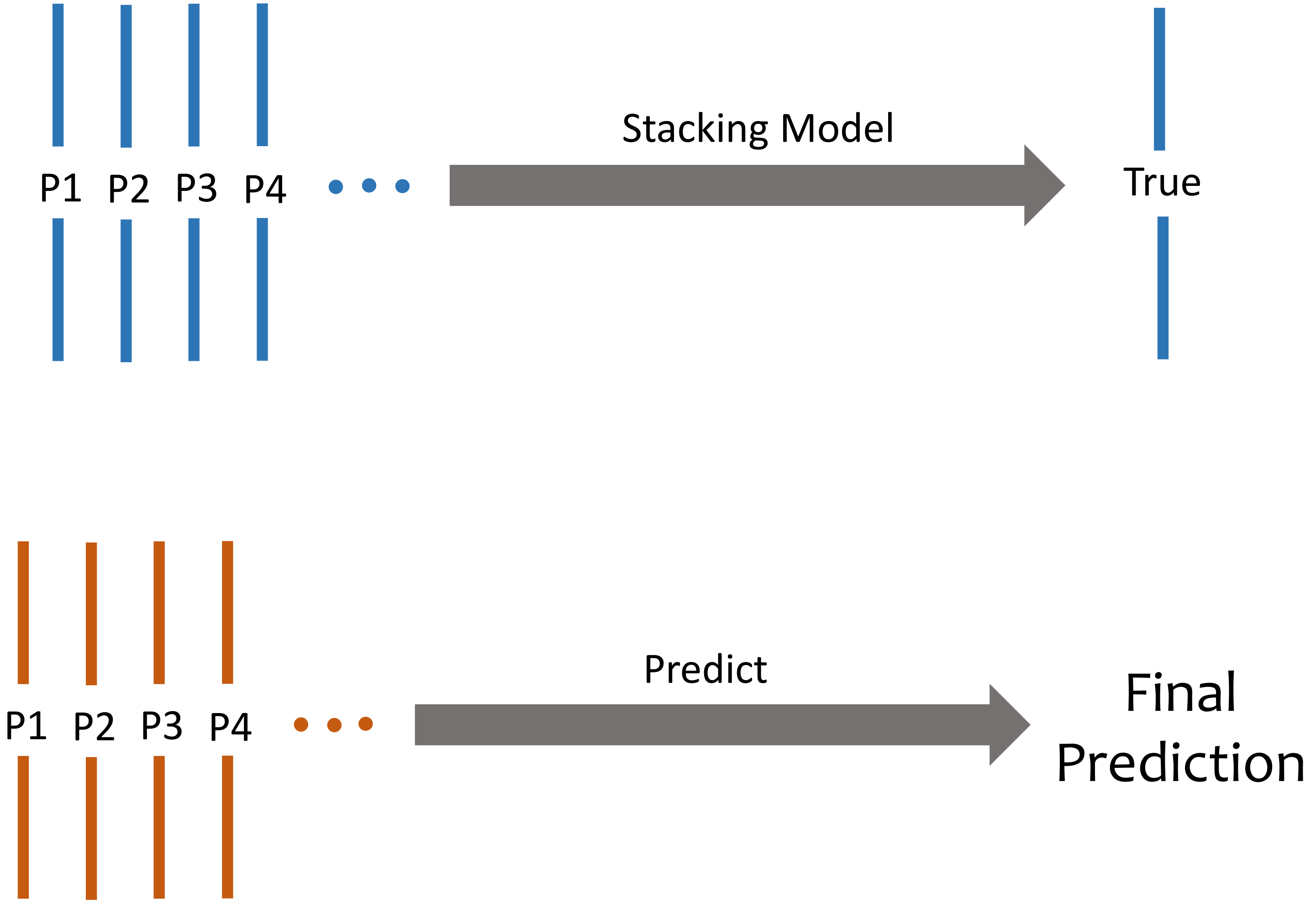
We get:



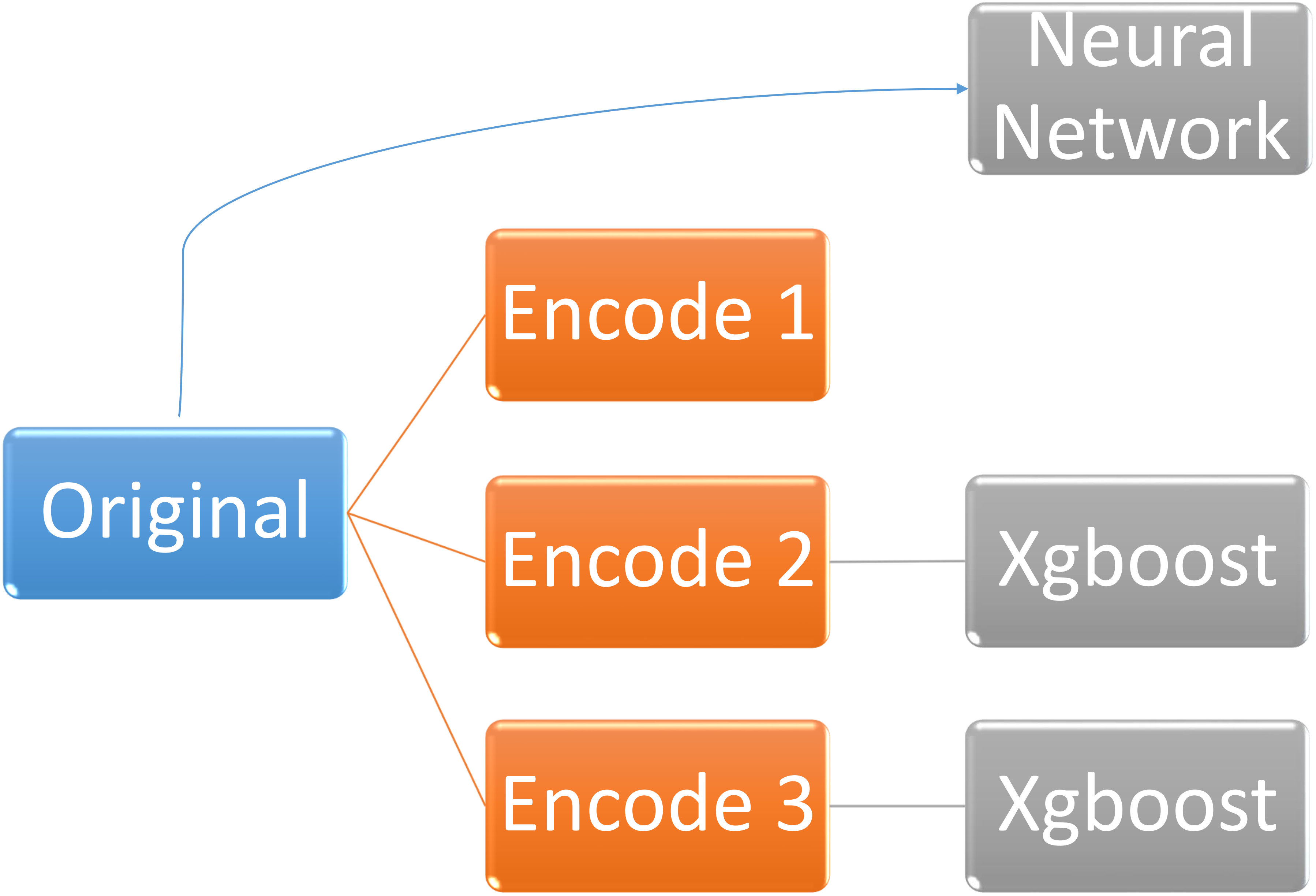
P1 P2 P3 P4

but no outcome
this time

# 5.1 Workflow for Stacking



# 5.2 Stacking Candidates



Neural Network

Encode 1

Original

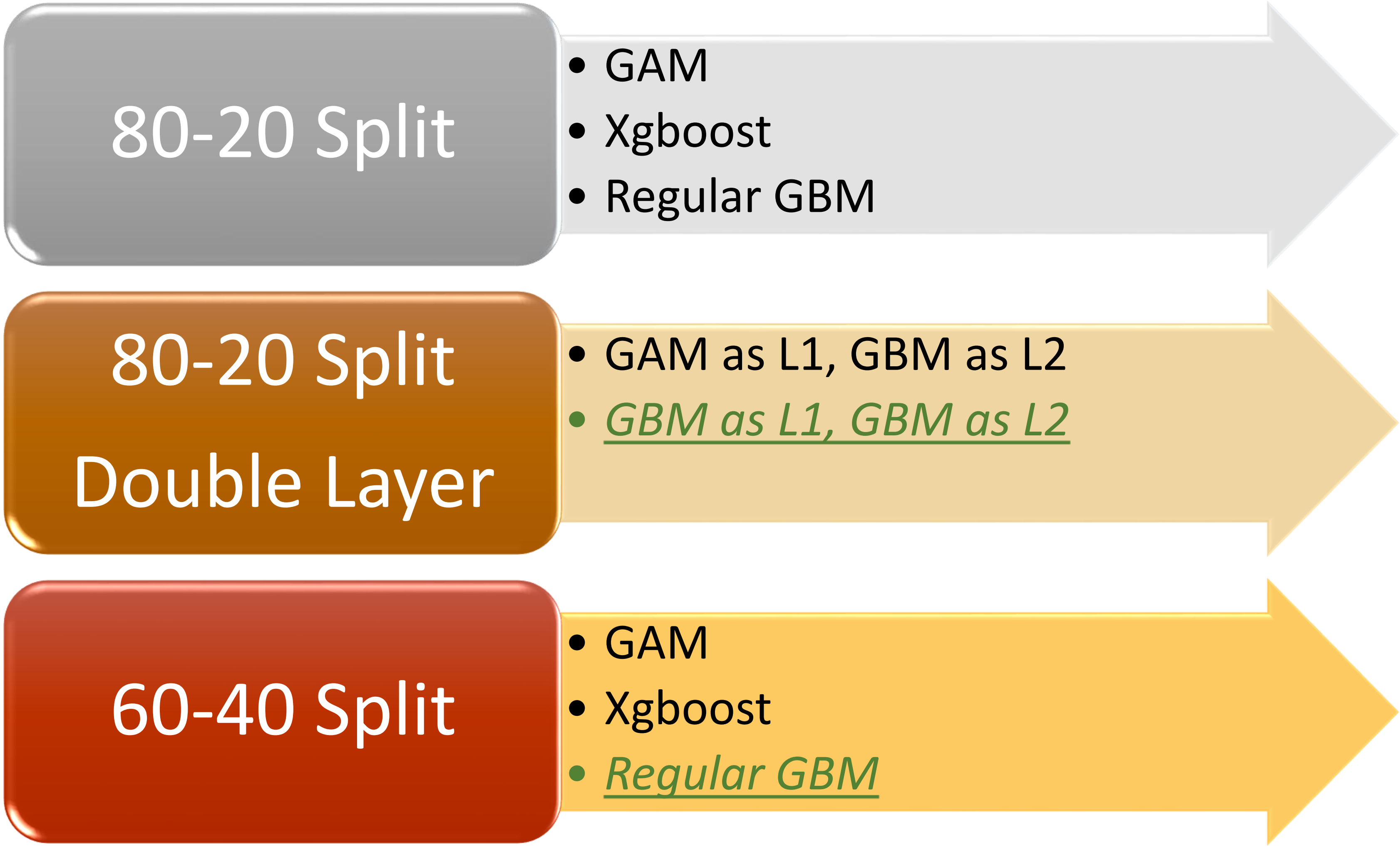
Encode 2

Xgboost

Encode 3

Xgboost

# 5.3 Stacking Attempts



80-20 Split

- GAM
- Xgboost
- Regular GBM

80-20 Split

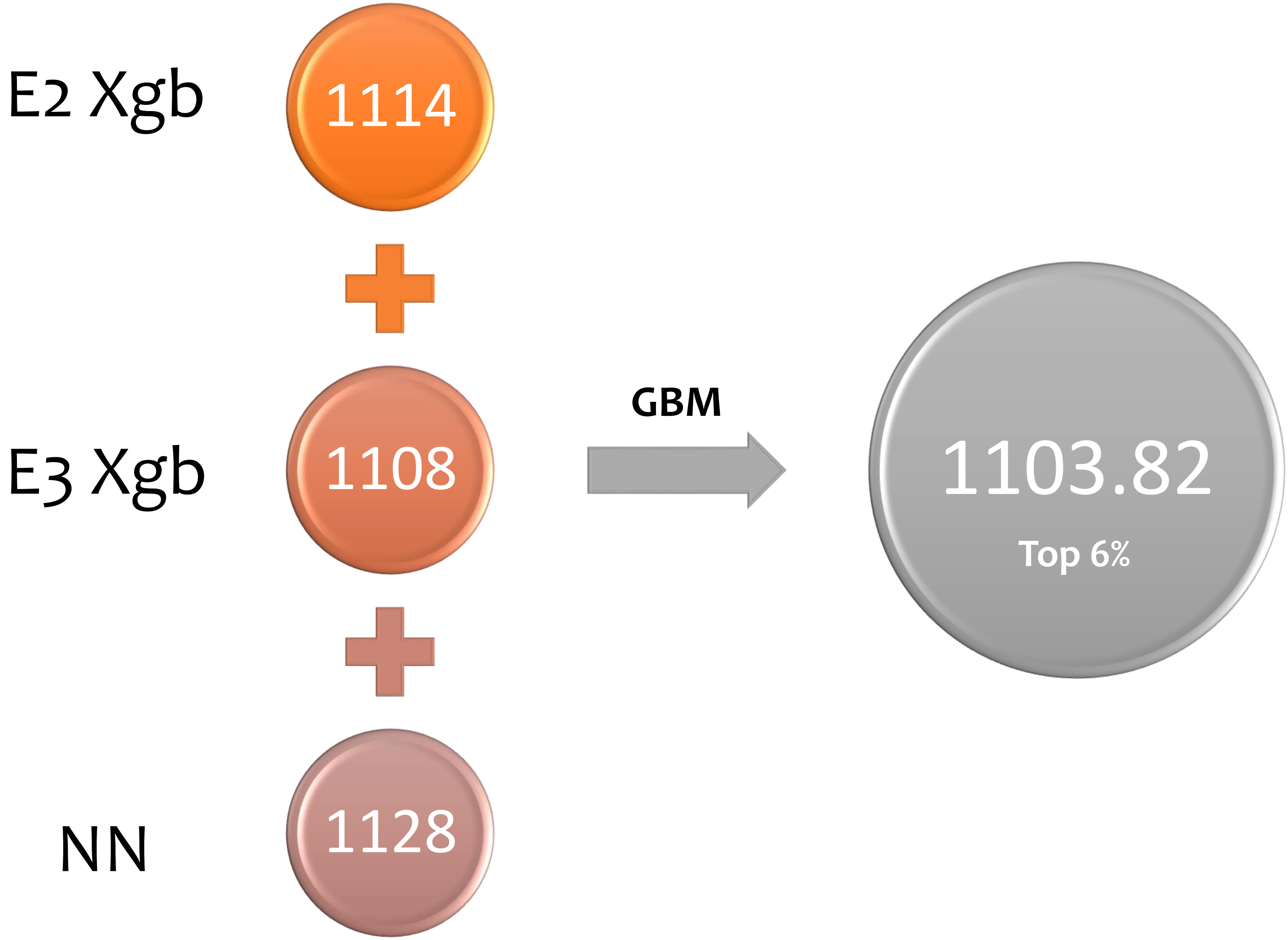
Double Layer

- GAM as L1, GBM as L2
- *GBM as L1, GBM as L2*

60-40 Split

- GAM
- Xgboost
- *Regular GBM*

# 5.4 Stacking Results



E2 Xgb 1114

E3 Xgb 1108

NN 1128

GBM

1103.82

Top 6%

## 6 Conclusion and Future Direction

1. Xgboost and Neural Network are two accurate algorithm for this dataset.

2. We were able to push the MAE to 1103.8 based on model stacking.

3. Multiple Neural Networks regarding to different encoding will be built to further climb the leaderboard.