


STAT 441 Project



A large yellow circle, resembling a sun or moon, is positioned behind the text 'STAT 441 Project'. The text is in a bold, black, serif font. Below the main title, there is a faint, grey, outlined version of the same text 'STAT 441 Project'.

Group 23/Team Struggle
Adyant Tian Jason

Table of Contents

01 Preprocessing

- Missing values
- One-hot encoding
- Normalization
- Null Values

02 Modeling

- XGBoosting
- Neural network
- Failed attempts

03 Stacking

04 Conclusion



01

Preprocessing



Data Processing

ID Data Type

Categorized numeric variables based on range to determine data type

Standardize Numeric Data

Numeric data was standardized to be in the range of $[-1,1]$
$$(x - \mu) / \sigma^2$$

One-Hot Encoding

Non-ordinal categorical data converted to binary values

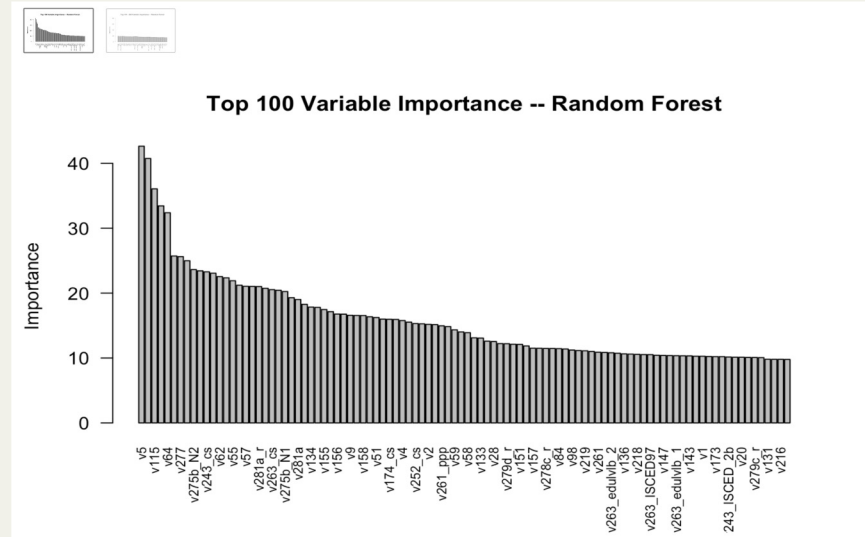
Null Values

10 observation among 4 variables with null values. Replaced with variables' rounded mean

Important Variables

Feature Selection

v63, v5, v54, v64, v115,
v56, v57, v282, v93,
v52_cs, v277,
v261_ppp



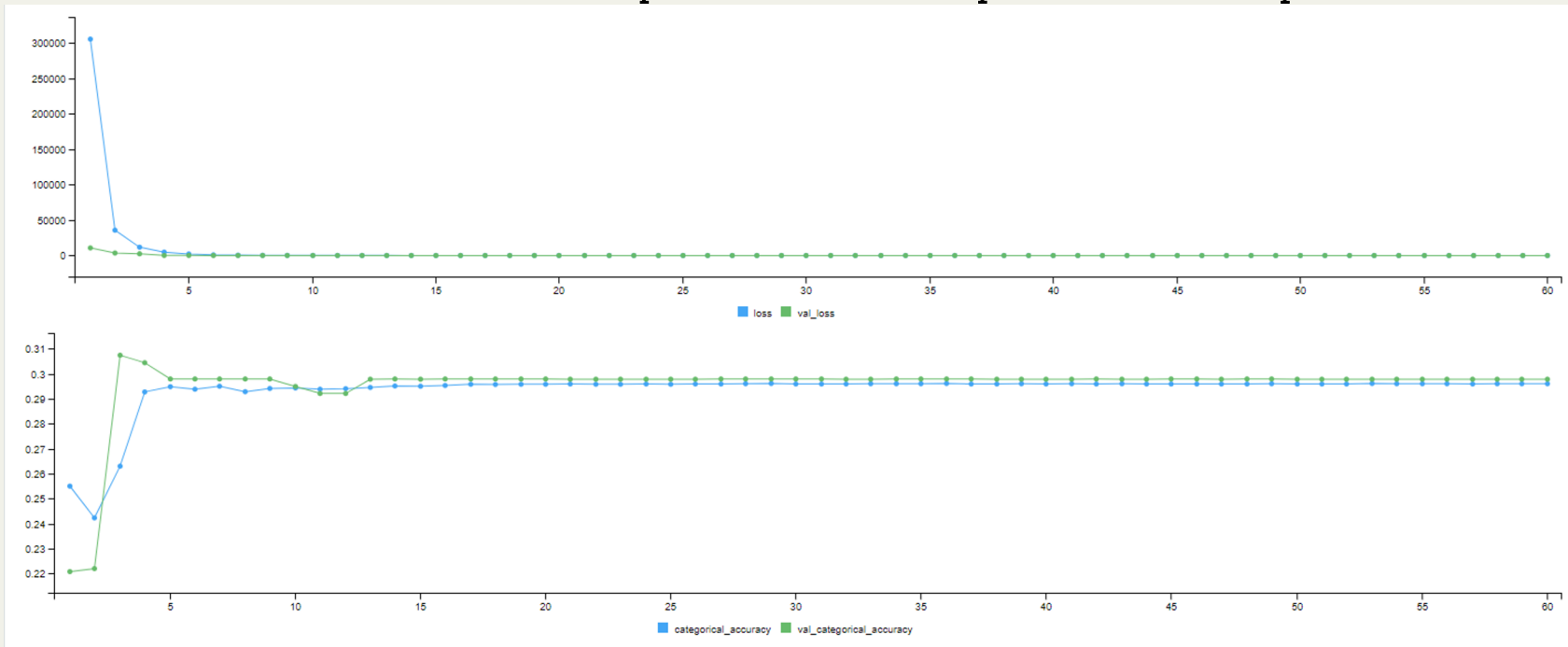
V63 - how important is God in your life
V5 - how important in your life: politics
V54 - how often attend religious services
V64 - how often do you pray outside religious services
V115 - how much confidence in: church



Modelling

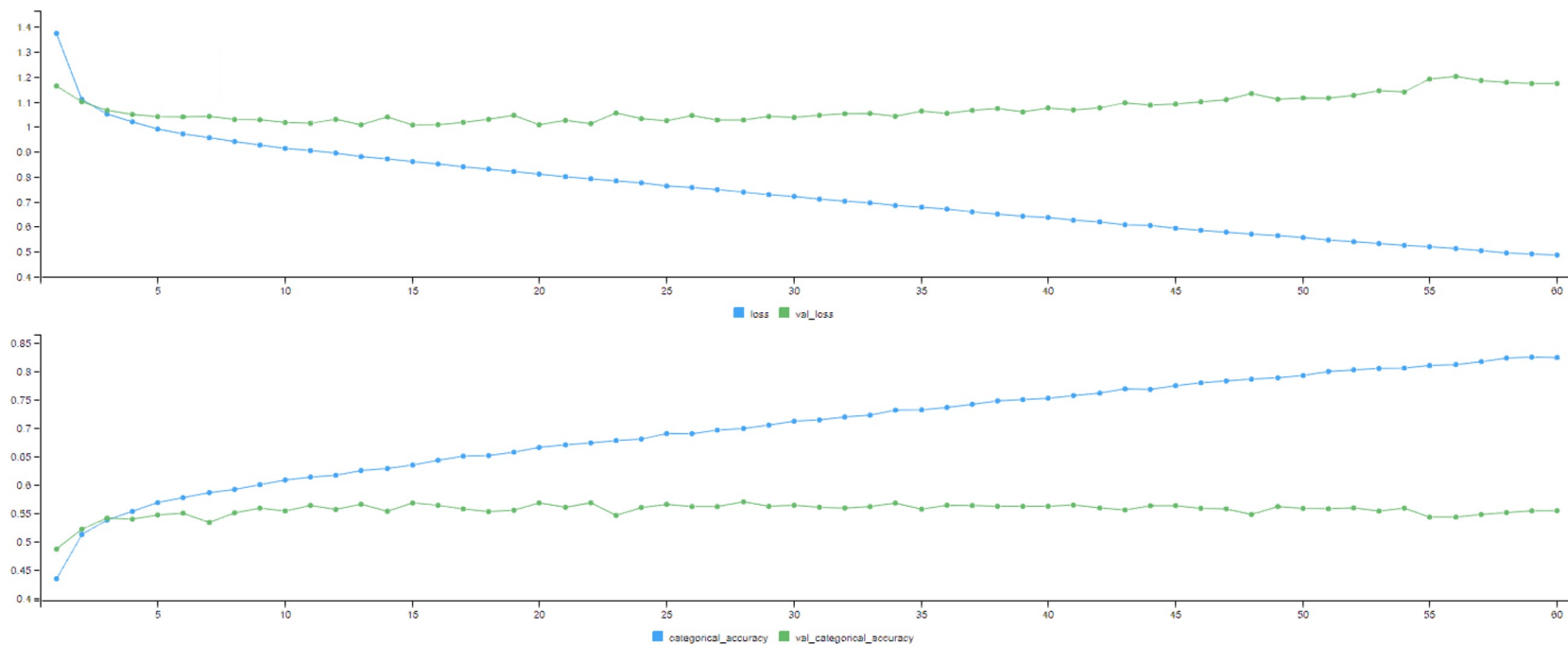
Neural Network - Attempt 1

Failed: Using the data without standardizing it. Due to data are measured in different scales. The results are not ideal, Caped at 0.3 Accuracy in both fit and predication



Neural Network - Attempt 2

Failed: Adding more layers does not improve the prediction power, but able the Neural Network to fit to the dataset very well. I.e. Overfit. However, by standardizing data, we

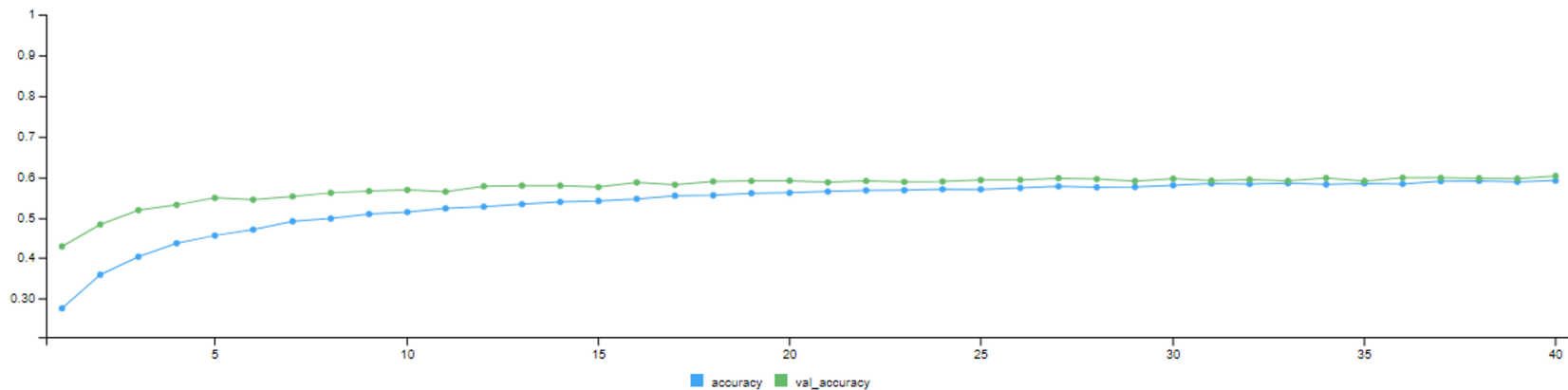
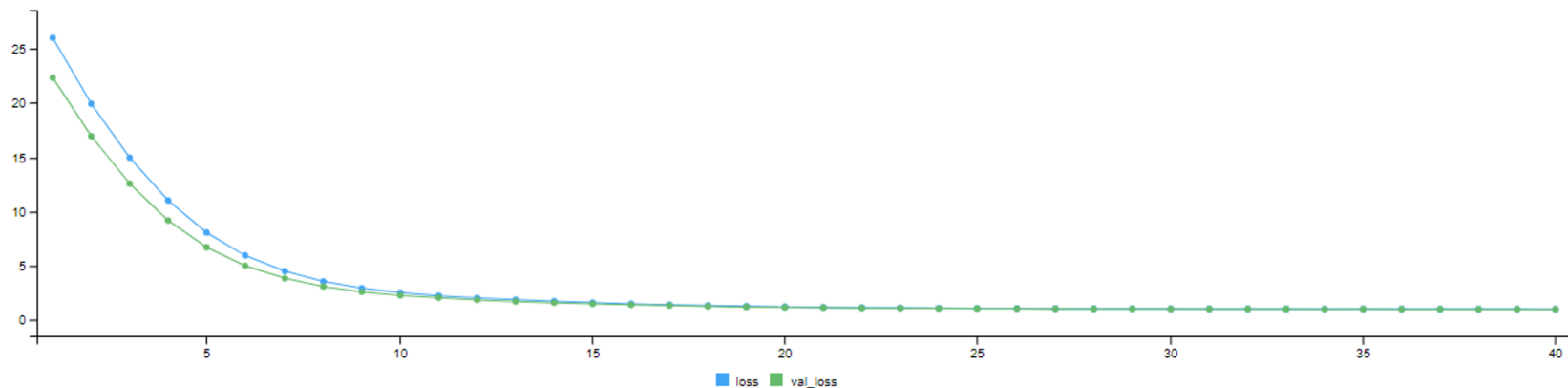


Neural Network - final

Final Neural Network:

- Two hidden layer,
 - 200 neurals, with L2 penalty
 - Relu as activation function
- Dropout Layer for Regularization technique to prevent overfitting
- Batch normalization layer to maintain mean close to 0 and standard deviation close to 1.
- Softmax as activation function for output layer

Neural Network - final Graph



XGBoost

- Reasons for choosing:
 - Higher performance
 - Reducing overfitting
- Parameter Fine Tuning
 - Grid Search
 - Cross Validation
- Final XGBoost Model:
 - Max depth = 5
 - Learning rate = 0.1
 - Iterations/N_Estimators= 250
- Attempted Feature Selection (Failed Attempt)

```
param_grid = {  
    'n_estimators': [100,200,400,600,800,1000],  
    'max_depth': [1,5,10,15,20],  
    'learning_rate': [0.001,0.01,0.1,1]  
}
```

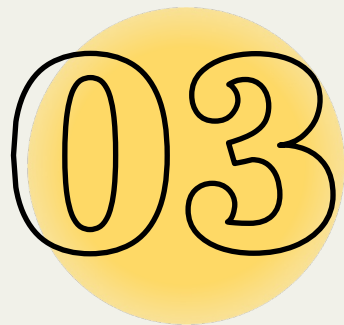
Initial grid search

:

```
param_grid = {  
    'n_estimators': [210,220,230,240,250,260,270,280,290],  
    'max_depth': [2,3,5,6,7,8,9],  
    'learning_rate': [0.09,0.1,0.11,0.12]  
}
```

Final grid search

Public Kaggle Score: 0.86852



Stacking


Stacking - attempt 1

Base Classifiers:

- Fine-tuned XGboosting (CV score: 0.643)
- Fine-tuned Random Forest (CV score: 0.603)
- Fine-tuned Neural Network (CV score: 0.612)

Final Classifier:

- Fine-tuned XGboosting



Stacking - attempt 1

Results: performed worse than XGboosting (XGboost public Kaggle score: 0.86852)

- Cross validation score: 0.641
- Kaggle score: 0.928 (public), 0.916 (private)

Concerns: Overfitting

- Tried using simpler classifier to prevent overfitting, which leads to our second attempt

Stacking - attempt 2

Base Classifiers:

- Fine-tuned XGboosting
- Fine-tuned Neural Network
- Fine-tuned Random Forest

Final Classifier:

- logistic

Cross validation score: 0.602

Stacking - Insights

- Aiming to fully leverage the advantages of stacking, we tried a few different stacking approaches.
- None of them performed better than the XGBoosting model
- Stacking doesn't provide any significant advantage for this particular dataset



Conclusion

Conclusion

- For large dataset like this (400+ predictors and 48,000 observations)
 - the model need to be computational efficient
 - preprocessing is the key (standardization, one-hot encoding, etc.)
- XGBoosting gives the best score

Did you know?

This project is known as laptop killer, effectively turn three laptop into three toaster with computing power

This image is generated by
DALL·E



Thank you