Credit Scoring Model Based on Financial and Relevant Profile Data

Saqib Al Islam - 16101084, Rifah Sama Aziz - 19141019, Aritra Ahmed - 16101216, Fauzia Abida - 16101320
Supervised By: Mahbub Alam Majumdar(PhD), Professor, Co-Supervised By: Md.Saiful Islam, Lecturer

Department of CSE, Brac University

Abstract

A credit score is a numerical expression based on a level analysis of an individual's credit files, to represent the creditworthiness of an individual. The credit score plays a major role in banks, financial institutions loaning money to individuals for their personal or business needs. This score is given based on factors such as personal information, assets, financial behavior and financial history. This system is not digitized or implemented yet in Bangladesh. So our aim is to build a reliable and robust credit scoring model which would help institutions like such to have an accurate reference score to rely on when validating a client. We were able to obtain an optimized model with an accuracy of (93%) on a Bank Loan Data-Set. The model is based on CART(Classification and Regression Trees) using Gradient Boosting method(GBM). We also proposed a new hybrid model consisting of a two step architecture(RfDNN). Since, credit scoring an individual is a sensitive issue, it is not ethical to treat the model as a 'Black-Box'. We conducted interpret-ability analysis on our model and generated visual representations of the criterion affecting the output of our model and provide necessary information to analyze the client effectively. Our results were conclusive and imitated the process of evaluating an individual precisely. The work-flow we proposed could be implemented in production to provide a concrete base for evaluation and prediction of defaulters. Simultaneously provide a detailed overview of the results obtained. This could help financial institutions immensely and help them save millions lost by default loans.

Keywords: Credit Score, Credit Risk, Loan Assessment, Machine Learning, Artificial Intelligence, Random Forests, Gradient Boosting, GBM, Extreme Gradient Boosting, KNN, RF, Deep Neural Networks, DNN, fDNN, Interpret-ability, LIME.

Literature Review

Credit risk assessment is a prominent topic in the field of banking and financing. Statistics and human evaluation are the key studies closely associated with it since it's instantiation. However recently due to the rapid advancements in data science and machine learning, credit risk assessment using Pattern recognition and Machine Learning have gained great significance in the research community. Plenty of noteworthy research papers have been published which gained traction in this field of study. These include Supervised Models, Logistic Regression, CART Tree-based Models, Neural Network models and ensemble methods along with their comparative analysis[1],[2], [3]

We found significant insight on the various difficulties and challenges associated with credit scoring from previous researches done on the topic.

Such as [4] where they used Generalized Linear model algorithm, which is a modification of logistic regression model. GLM is an improvement of binary classification as it provides a confidence bound where lies the probability of a positive outcome.

Furthur improvements were proposed by [5] who used Ensemble Logistic Regression boosted by GradientBoost[6]

Taking into account all the findings we have decided to use an ensemble model approach to our given problem.

References

[1] Peter Addo, Dominique Guegan, and Bertrand Hassani.
Credit risk analysis using machine and deep learning models.
Risks, 6(2):38, 2018.

[2] Bhekisipho Twala.
Multiple classifier application to credit risk assessment.
Expert Systems with Applications, 37(4):3326âĂŞ3336, 2010.

[3] L Yu, S Wang, and K Lai.
Credit risk assessment with a multistage neural network ensemble learning approach.
Expert Systems with Applications, 34(2):1434âĂŞ1444, 2008.

[4] Jasmina Nalić and Amar Švraka.
Using data mining approaches to build credit scoring model: Case studyâĂŤimplementation of credit scoring model in microfinance institution.
In *INFOTEH-JAHORINA (INFOTEH), 2018 17th International Symposium*, pages 1–5. IEEE, 2018.

[5] A. Lawi, F. Aziz, and S. Syarif.
Ensemble gradientboost for increasing classification accuracy of credit scoring.
In *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*, pages 1–4, Aug 2017.

[6] Pradeep Singh.
Comparative study of individual and ensemble methods of classification for credit scoring.
In *Inventive Computing and Informatics (ICICI), International Conference on*, pages 968–972. IEEE, 2017.

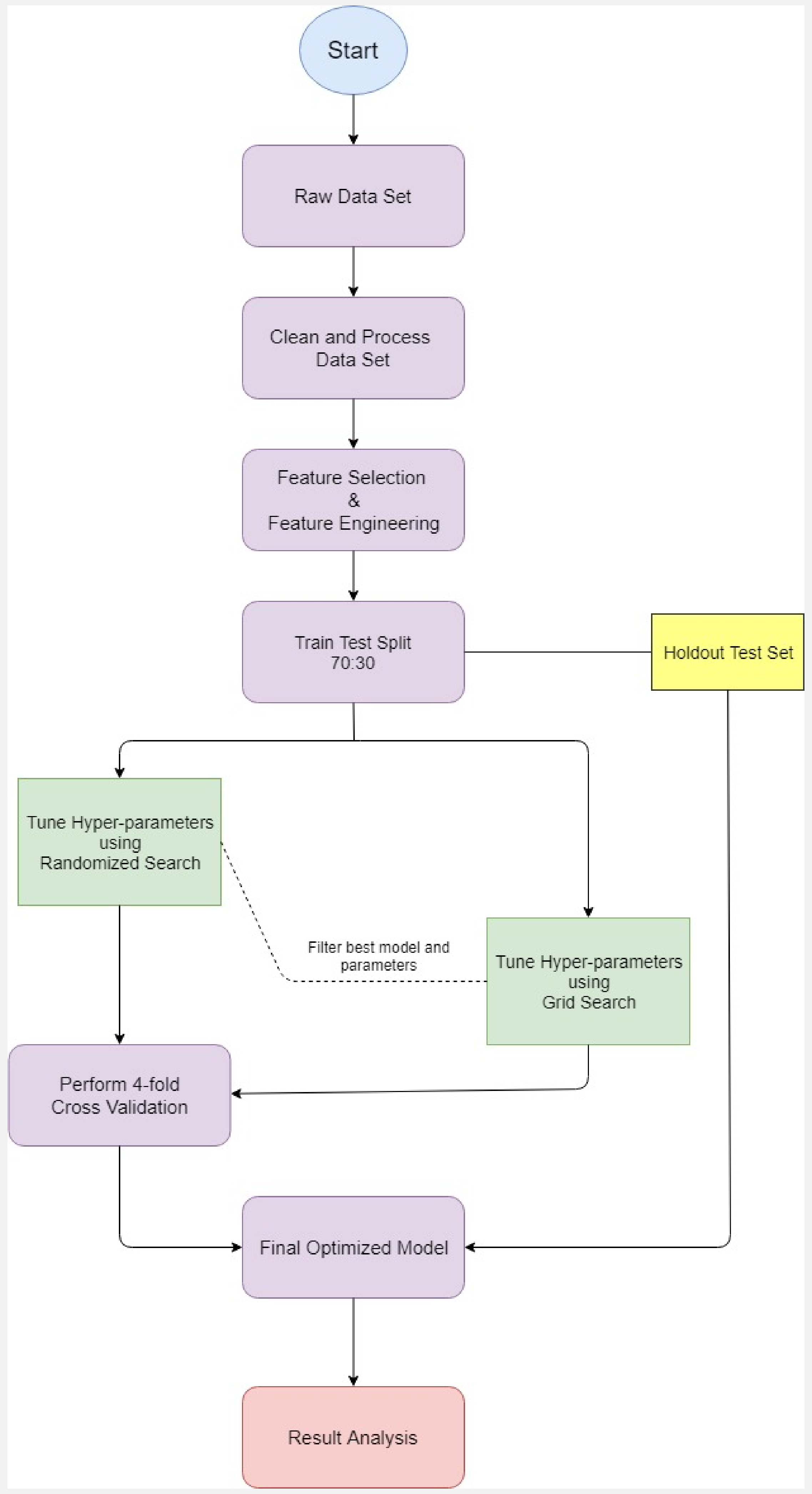

Start
Raw Data Set
Clean and Process Data Set
Feature Selection & Feature Engineering
Train Test Split 70:30
Holdout Test Set
Tune Hyper-parameters using Randomized Search
Filter best model and parameters
Tune Hyper-parameters using Grid Search
Perform 4-fold Cross Validation
Final Optimized Model
Result Analysis

Figure: Overview of Workflow

Work-Flow

Following is an overview of our work-flow:

- Raw Data-set is Cleaned and Processed. This includes imputing missing values, identifying outliers and feature scaling
- Feature Selection and Engineering is carried out. Selection of important features using correlation and feature importance as reference.
- Train Test split carried out in the ratio 70:30, where the test set is held out for evaluation in final model.
- Randomized Search over hyper-parameters of selected models are conducted to further filter models and parameters to conduct Grid Search on.
- Models are optimized using Cross-Validation along with Randomized and Grid Search over the hyper-parameter space
- Final model evaluated on hold out test set. Results obtained and analyzed. Comparative analysis conducted between the models and the final model.

Proposed RfDNN model

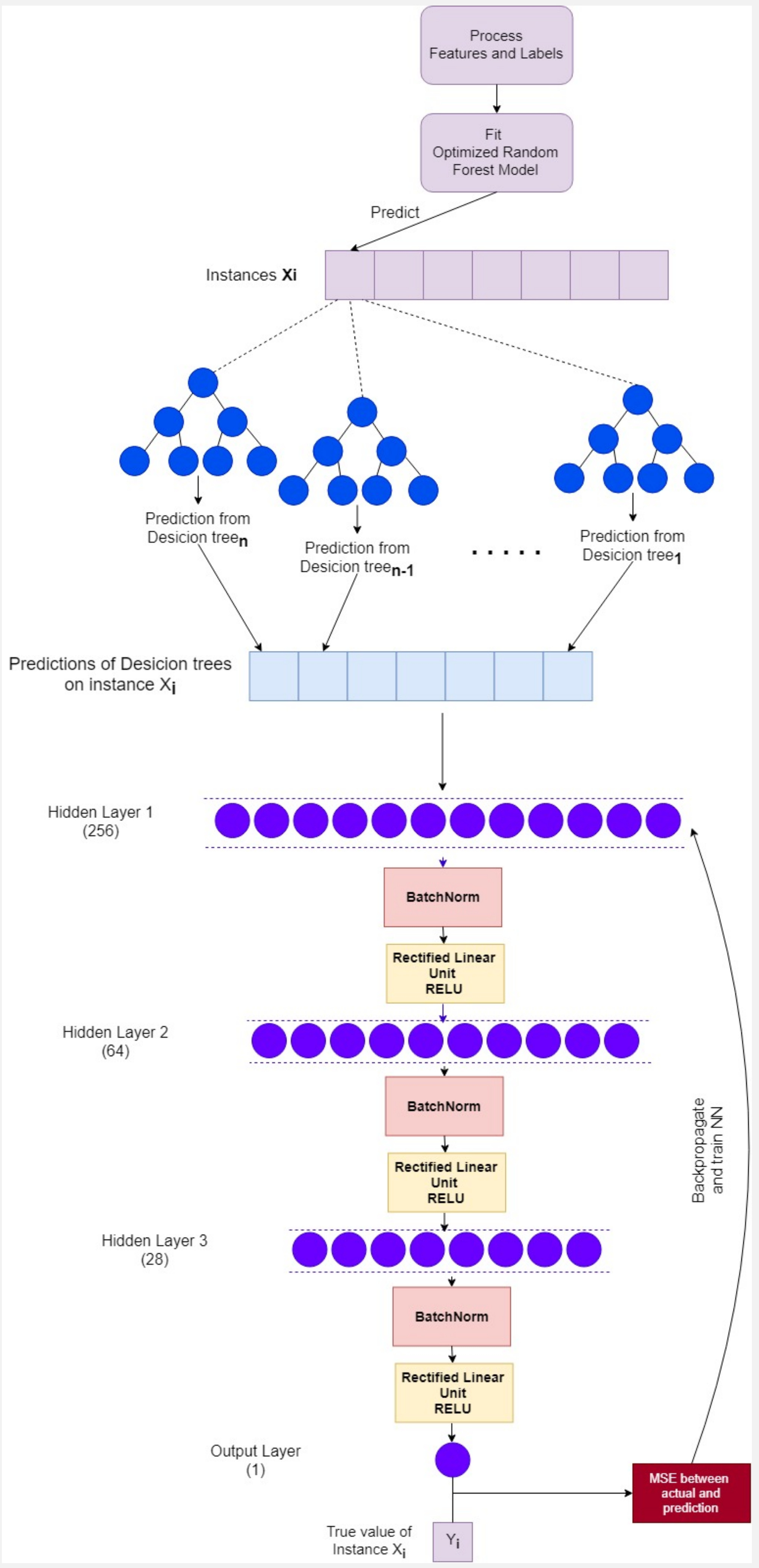
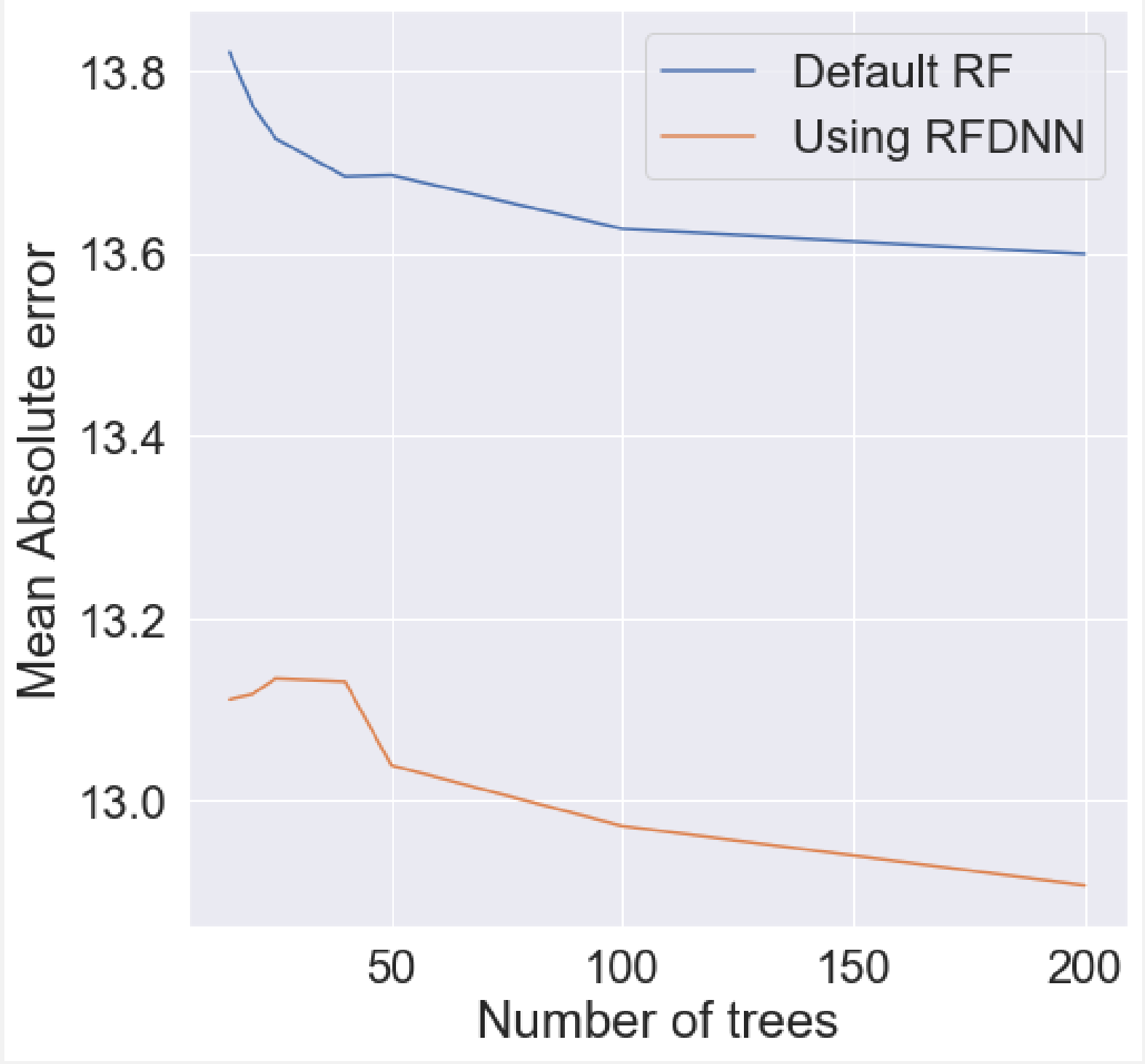

Process Features and Labels
Fit Optimized Random Forest Model
Predict
Instances **Xi**
Prediction from Desicion tree$_n$
Prediction from Desicion tree$_{n-1}$
. . . . .
Prediction from Desicion tree$_1$
Predictions of Desicion trees on instance X_i
Hidden Layer 1 (256)
BatchNorm
Rectified Linear Unit RELU
Hidden Layer 2 (64)
BatchNorm
Rectified Linear Unit RELU
Hidden Layer 3 (28)
BatchNorm
Rectified Linear Unit RELU
Output Layer (1)
MSE between actual and prediction
Backpropagate and train NN
True value of Instance X_i
Y_i

Figure: Proposed RfDNN model Architecture

RfDNN Result Analysis

After two step training of the RfDNN model for 20 epochs with mini-batch of 10,000 instances, it was tested against Optimized Random Forest with different tree numbers. The results obtained are as follows:



Experimental Results and Analysis

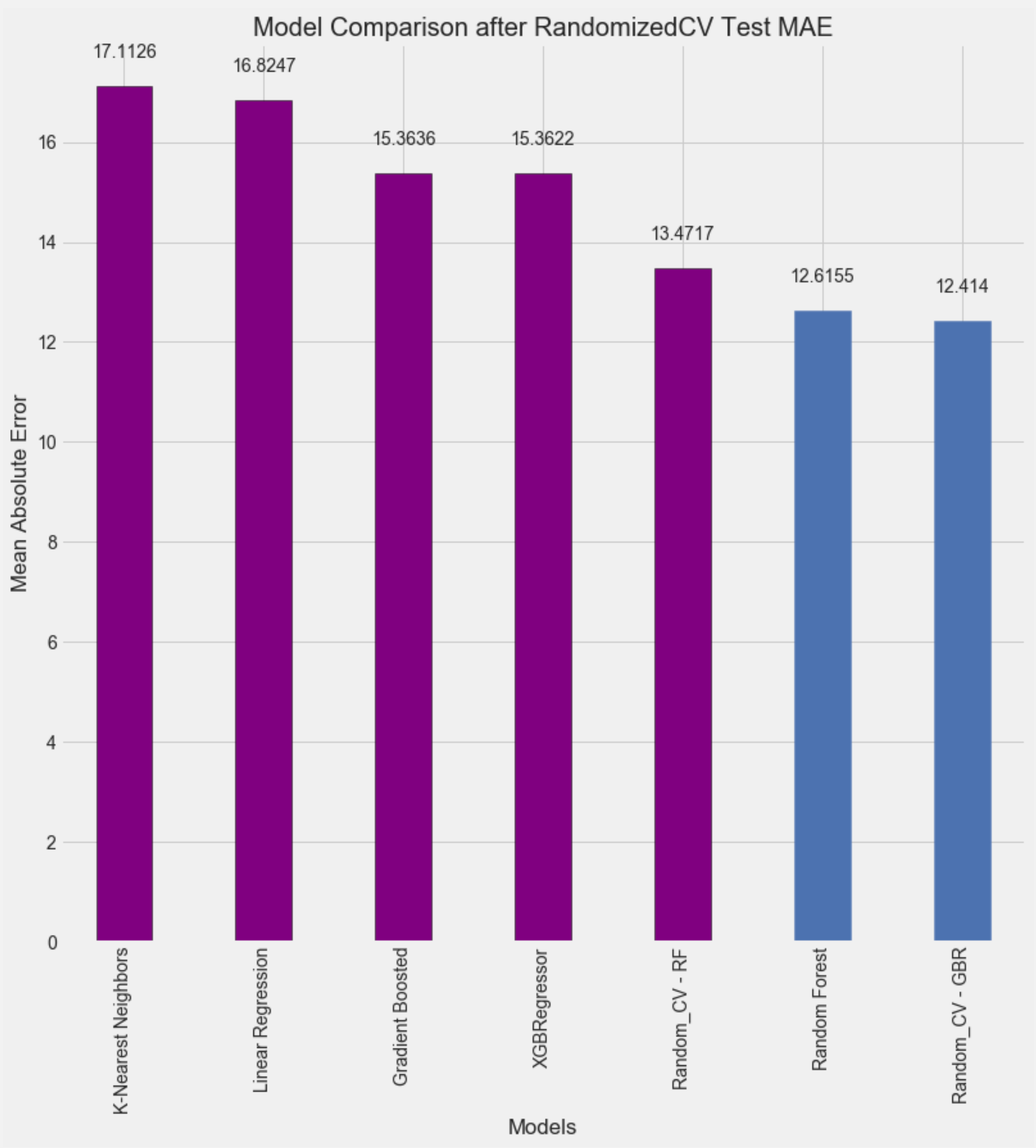


Figure: Comparative Analysis of models

Model	MAE	accuracy
GridCV - GBR	11.57	93.02%
RandomCV - GBR	12.413	92.52%
RandomCV - RF	12.615	92.40%
Default Random Forest	13.471	91.88%
XGBRegressor	15.362	90.76%
Default GBR	15.363	90.74%
Linear Regression	16.824	89.86%
K-Nearest Neighbors	17.112	89.69%

Table: Accuracy of Different models

Although the final optimized GBR model was good at predicting scores below the eligibility threshold. It was not so accurate in predicting the bi-modal distribution of credit score above the eligibility threshold. That is not a major concern because the purpose of credit scoring is to accurately predict the defaulters and ineligibles.

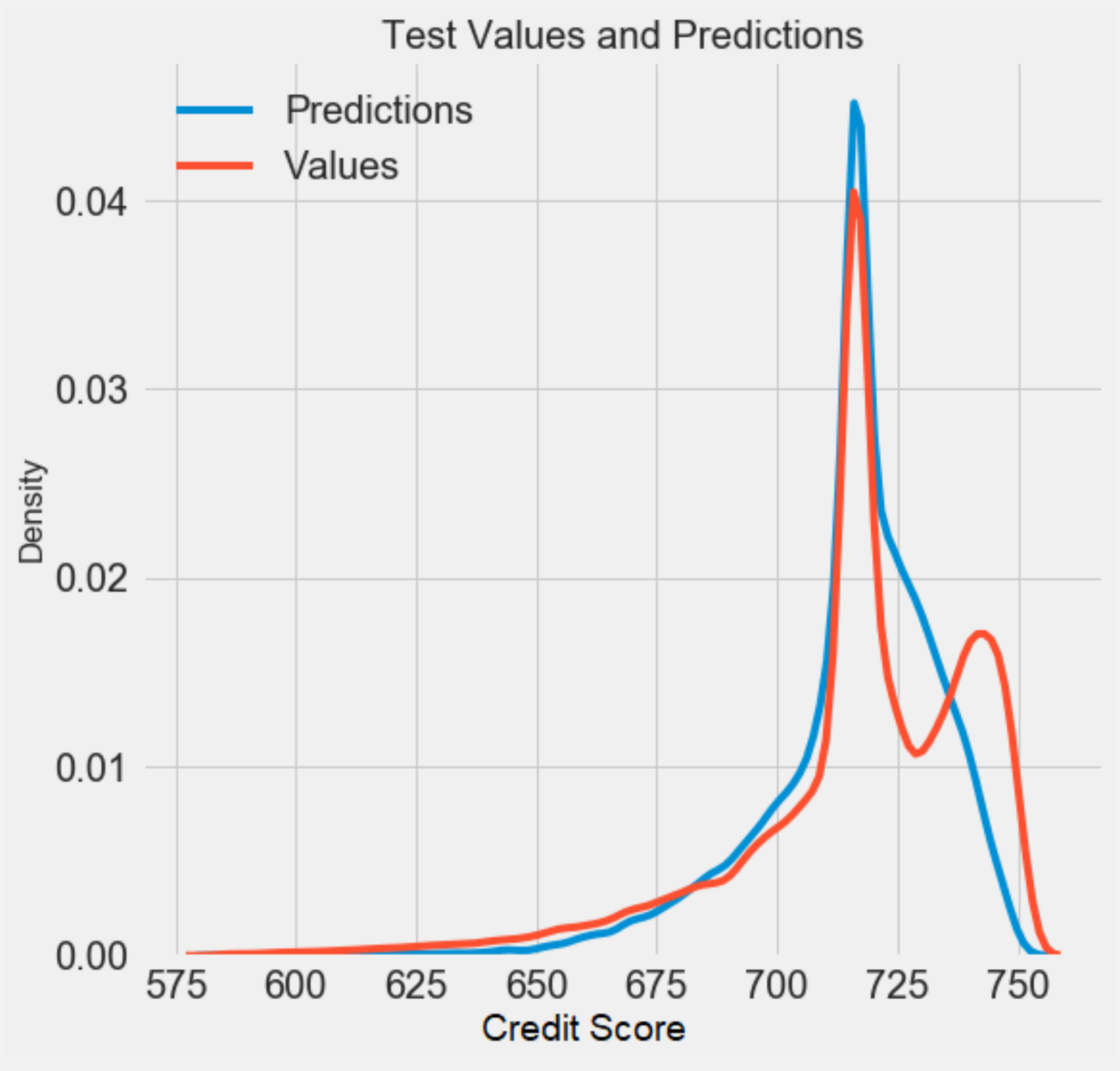


Figure: Predictions and True Value Distribution of Test Set

Conclusion & Future Work

We were able to come up with two models achieving high accuracy on the test set. We plan to work on reducing the computational cost of training the RfDNN model. We would also look forward to testing our trained model on real-world data to determine its accuracy.