CREDIT RISK ASSESSMENT FOR HOME CREDIT GROUP

K.V.Shashank[1], Vivek Gutti[2]

[1]Post Graduate Student, Department of Computer Science & Engineering, Amrita University-Coimbatore, Tamil Nadu, India
[2]Post Graduate Student, Department of Computer Science & Engineering, Amrita University-Coimbatore, Tamil Nadu, India

---***---

Abstract – *Home Credit could be a world non-Banking institution (NBFC) based mostly in 1997 at intervals the US nation. the company operates in fourteen countries and focuses on disposal primarily to parents with little or no or no credit history. There unit of measurement 307511 observations at intervals the dataset with 122 columns that represent every qualitative and quantitative attribute of those sixty-seven columns have missing values. Out of 122, sixteen columns unit of measurement categorical and 106 unit of measurement numerical columns. there is a binary output variable that denotes "Delay in payments" (1) or "No Delay in payments (0)".*

Key Words: Non-banking, Institution, delay, no delay in payments, Home Credit, binary output.

1.INTRODUCTION

Home Credit strives to broaden cash inclusion for the unbanked population by providing a positive and safe borrowing experience. therefore, on kind bound this underserved population options a positive loan experience, Home Credit makes use of a variety of various data--including service and transactional information--to predict their clients' compensation skills. Home Credit is presently victimization varied applied math and machine learning ways in which to make these predictions successful. This could check that the purchaser is capable of compensation are not rejected that loans unit of measurement given with a principal, maturity, and compensation calendar that will empower their purchasers to attain success.

1.1 PROJECT JUSTIFICATION

Conservative credit risk management policies, fast loan decisions, and cheap loan valuation win this balance of protecting loan portfolios whereas keeping bank customers pleased with the institution.

The objective of this project is to predict the house loan credit risk for the institution. The project will modify the bank to chop back its risk of loan loss by gaining Associate in Nursing apt understanding of its consumer base, therefore minimizing the loss of capital for the financial institution whereas reaping best profit.

By analyzing the consumer choices like dealing history, annual gain, demographics etc., and thus the bank square measure attending to be able to estimate the danger of the loan compensation.

2. ALGORITHMS USED

2.1. Logistic Regression

The most common use of Logistic regression models is in binary classification problems. Logistic Regression is also a supervised classification model. It permits you to make predictions from labelled information if the target (output) variable is categorical. Used as a result of having a categorical outcome variable violates the thought of spatiality in ancient regression. instead of building a mantic model for "Y (Response)" directly, the approach models "Log Odds (Y)"; thence the name provision or Logic. the foremost draw back with a straight line is that it isn't steep enough. at intervals the sigmoid curve, as you will see, you have low values for various points, then the values rise all of a pointy, once that you have got various high values.

Sigmoid function operates as: - s(x)=1/1e^-x Log(odds)=log(p/1-p)z=beta0+beta1 h(x)=sigmoid(z) h(x)1/1+e^-(beta0+beta1)

2.2. Decision Trees

A decision tree uses a tree-like model to make predictions. It resembles Associate in Nursing turned tree. it's to boot really reasonably like but you produce decisions in real life, you raise a series of inquiries to achieve a decision. A decision tree splits the information into multiple sets. Then, each of these sets is further split into subsets to achieve a decision. The topmost decision node throughout a tree corresponds to the foremost effective predictor referred to as the idea node. A node whereas not further branches is termed a leaf node. The leaf nodes represent the last word decisions. we have a tendency to square measure able to calculate that node is that the foundation by looking for the information gain, entropy, or Gini index. It suffers from high bias and high variance and will be a greedy learner.

Entropy E(s)= Σ-p(xi) log2 p(xi), Range is 0 to 1 lesser the score its best used in (ID3, C4.5, C5.0) Gini Index=1- Σp(xi)square, Range is 0 to 0.5 lesser the score its best used in (CART) Information Gain=H(s)-

Σ|v|/|s|(H(V), It should be high less entropy or Gini index information gain or vice versa.

2.3. Random Forest

Random Forest is used for Associate in Nursing ensemble of decision trees. It uses the lowest principle of material with random feature option to kind a great deal of diverse trees. cacophonic a node throughout the event of a tree, the split that is chosen is not any longer the foremost effective split among all the choices. Instead, the split picked is that the most effective split among a random set of the choices. As a result of this randomness, the bias of the forest generally slightly can increase (with respect to the bias of 1 systematic tree). Random Forest will use sqrt root of n choices for classification and n/3 choices for regression whereas n being an entire vary of choices.

2.4. Bagging

Bagging stands for Bootstrap Aggregation. Bootstrapping suggests that creating bootstrap samples from a given information set. A bootstrap sample is created by sampling the given information set uniformly and with replacement. A bootstrap sample sometimes contains regarding 30-70% information from the information set. it is a parallel technique that means employment and testing square measure attending to be done parallelly and freelance of each totally different. fabric handles over-fitting and reduces variance.

2.5. Boosting

Boosting is also a sequent technique, where each ensuant model tries to correct the errors of the previous model. The succeeding model's unit of measurement obsessed to the previous model. Boosting provides misclassified samples higher weight. it is a thanks to boost weak learning algorithms (single tree) into a sturdy learning algorithm. employment and Testing unit of measurement sequent in Boosting. the aim of Boosting is to chop back Bias. it's attending to increase the overfitting at intervals the information.

2.6. Ada Boost

AdaBoost (Adaptive Boosting) works on up the areas wherever the bottom learner fails. the bottom learner could be a machine learning rule that's a weak learner and upon that the boosting technique is applied to show it into a powerful learner. Any machine learning rule that accepts weights on coaching knowledge may be used as a base learner. AdaBoost works on up the areas wherever the bottom learner fails. the bottom learner could be a machine learning rule that's a weak learner and upon that the boosting technique is applied to show it into a powerful learner.

2.7. Gradient Boosting

Gradient boosting doesn't modify the statistical distribution. rather than coaching on a brand-new statistical distribution, the weak learner trains on the remaining errors (so-called pseudo residuals) of the sturdy learner. it's in a different way to grant additional importance to troublesome instances. At every iteration, the residuals are computed and a weak learner is fitted to those residuals.

2.8. XGBoost

XGBoost (Extreme Gradient Boosting) is associate optimized distributed gradient boosting library. It uses a gradient boosting (GBM) framework at the core. Yet, will higher than the GBM framework alone. XGBoost implements data processing and is quicker as compared to GBM. XGBoost has associate in-built routine to handle missing values. XGBoost tries various things because it encounters a missing worth on every node and learns that path to require for missing values in future.

3. EXPERIMENTAL ANALYSIS

3.1 Class Imbalance

Oversampling the target variable by using SMOTE. Classification using category-imbalanced knowledge is biased in favor of the bulk class. Oversampling is that the method of changing the minority category into the bulk category. i.e., increase the minority category to majority category count. The artificial Minority Over-sampling Technique (SMOTE) is associate oversampling approach that makes artificial minority category samples. It doubtless performs higher than straightforward oversampling and it's wide used.

3.2 Statistical Analysis

Statistical tests were performed to envision whether or not the independent variables have a big relationship with the variable quantity, TARGET.

3.2.1 Chi-square Test

For the explicit Columns, a Chi-square take a look at of independence was performed with the target variable, TARGET that is additionally a categorical column.

3.2.2 Two-sample t test

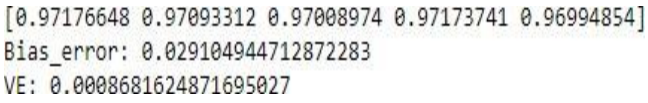
For all the numeric variables, two-sample mismatched t-tests were performed between values of the variable for 2 categories of target variables to match their means.

3.3 Min-Max Scalar

Min-max pulse counter for non-normalized options, get dummies for all options, and have integration.

3.4.1 PCA

Since the information is big, PCA is employed to envision whether or not this could improve our model performance. From the higher than results, it's ascertained that Accuracy, precision, recall, and f1-score are inflated when put next with the bottom model. when acting Cross-Validation for the PCA model with CV=5 and marking = "roc_auc‟, below are the results



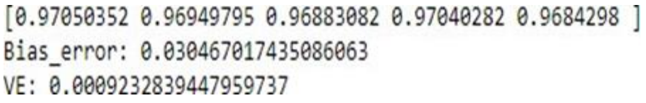
```
[0.97176648 0.97093312 0.97008974 0.97173741 0.96994854]
Bias_error: 0.029104944712872283
VE: 0.0008681624871695027
```

Though we have a tendency to get smart results, we wish to undertake the "Select KBEST" technique for feature choice and proceed with the model building as a result of when implementing PCA on the dataset, our original options can develop into Principal parts. Principal parts remain the linear combination of our original options. Principal parts don't seem to be as clear and explicable as original options.

3.5 Select K -Best

Statistical tests like Chi-square and t-test for the dataset, the results of the take a look at shown that everyone the options are important with relevance the target variable. within the choose K-Best technique, we have a tendency to specify the quantity of options and also the technique returns the foremost important amongst them.

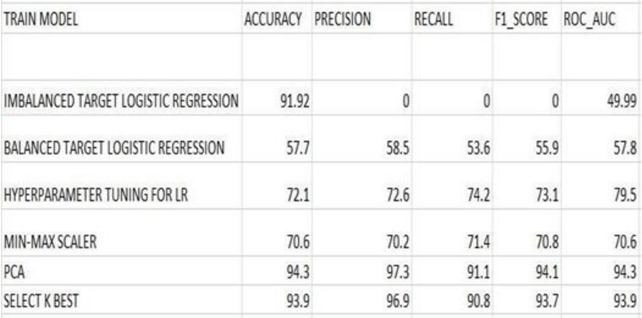
We had number of trials with k=80, 90,100 and so on. We got good score for k=100 value. Although the scores will be more if we go beyond k=120, but there would be much variance error in the scores. So, choosing optimal k value can be done only through trial-and-error method. After performing Cross Validation for the model with CV=5 and scoring = "roc_auc" , below are the results.



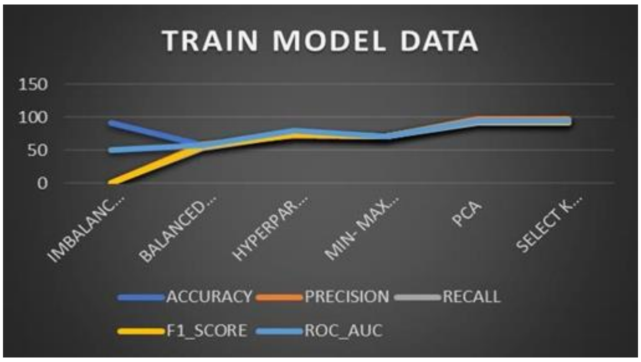
```
[0.97050352 0.96949795 0.96883082 0.97040282 0.9684298 ]
Bias_error: 0.030467017435086063
VE: 0.0009232839447959737
```

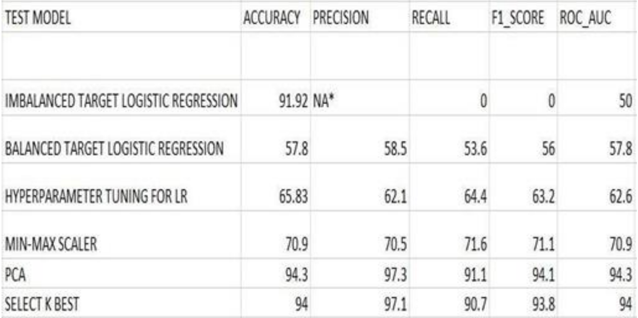
As our score are almost same as PCA, here we are able to interpret the features by backing them with statistical analysis. So going further we tried different ensemble methods on K Best Model to see if we can improve the score.

4. TABLES AND GRAPHS



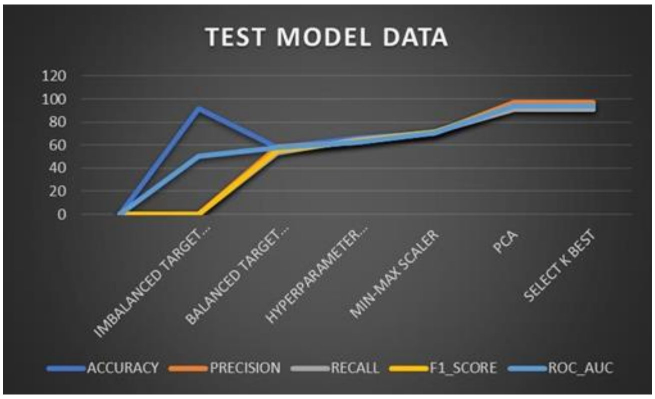
TRAIN MODEL	ACCURACY	PRECISION	RECALL	F1_SCORE	ROC_AUC
IMBALANCED TARGET LOGISTIC REGRESSION	91.92	0	0	0	49.99
BALANCED TARGET LOGISTIC REGRESSION	57.7	58.5	53.6	55.9	57.8
HYPERPARAMETER TUNING FOR LR	72.1	72.6	74.2	73.1	79.5
MIN-MAX SCALER	70.6	70.2	71.4	70.8	70.6
PCA	94.3	97.3	91.1	94.1	94.3
SELECT K BEST	93.9	96.9	90.8	93.7	93.9

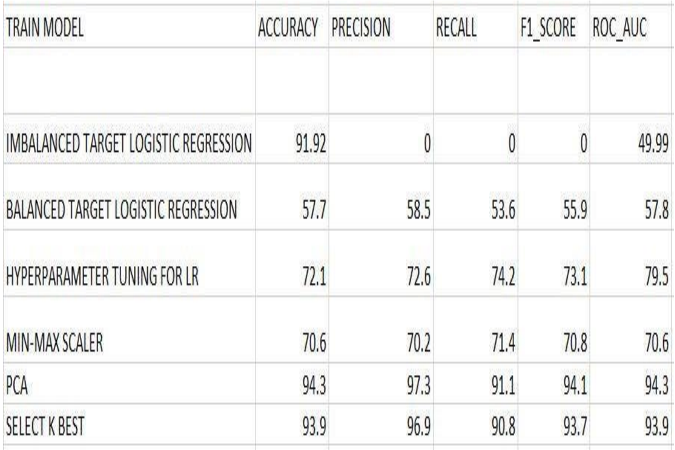




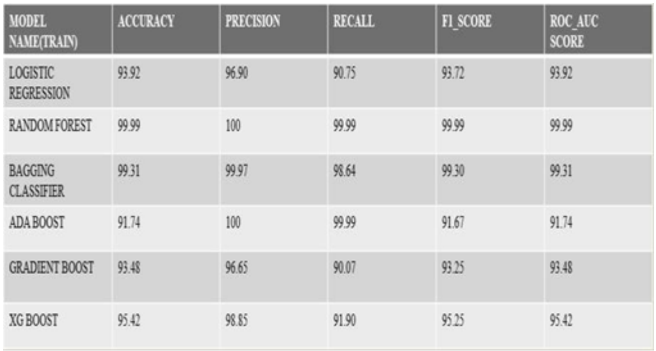
TEST MODEL	ACCURACY	PRECISION	RECALL	F1_SCORE	ROC_AUC
IMBALANCED TARGET LOGISTIC REGRESSION	91.92	NA*	0	0	50
BALANCED TARGET LOGISTIC REGRESSION	57.8	58.5	53.6	56	57.8
HYPERPARAMETER TUNING FOR LR	65.83	62.1	64.4	63.2	62.6
MIN-MAX SCALER	70.9	70.5	71.6	71.1	70.9
PCA	94.3	97.3	91.1	94.1	94.3
SELECT K BEST	94	97.1	90.7	93.8	94

NA* : - It is an edge case, model hasn't predicted any positive cases due to class imbalance

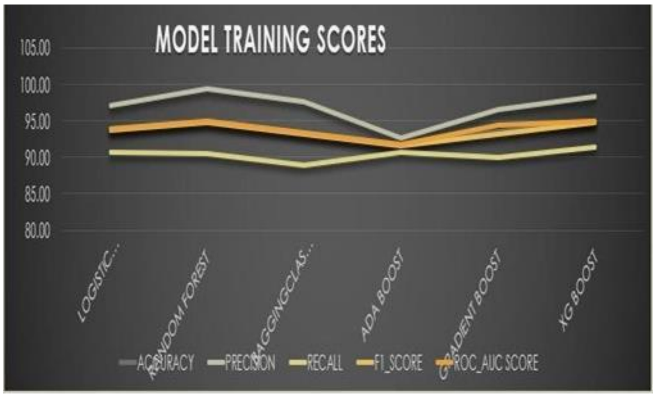


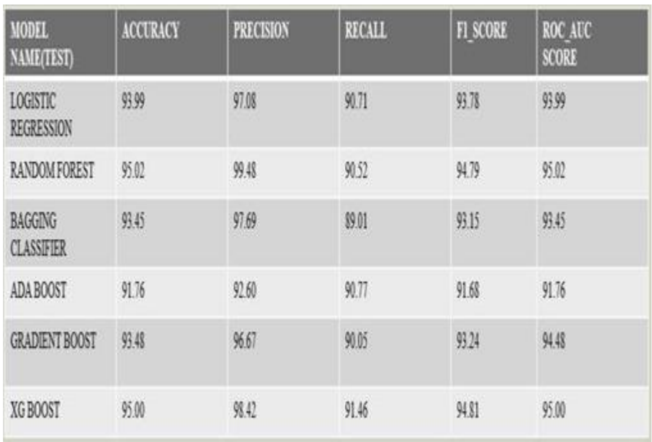


TRAIN MODEL	ACCURACY	PRECISION	RECALL	F1_SCORE	ROC_AUC
IMBALANCED TARGET LOGISTIC REGRESSION	91.92	0	0	0	49.99
BALANCED TARGET LOGISTIC REGRESSION	57.7	58.5	53.6	55.9	57.8
HYPERPARAMETER TUNING FOR LR	72.1	72.6	74.2	73.1	79.5
MIN-MAX SCALER	70.6	70.2	71.4	70.8	70.6
PCA	94.3	97.3	91.1	94.1	94.3
SELECT K BEST	93.9	96.9	90.8	93.7	93.9

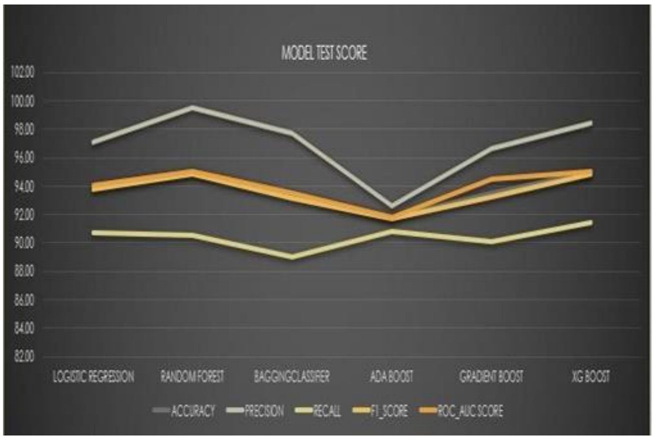


MODEL NAME(TRAIN)	ACCURACY	PRECISION	RECALL	FI_SCORE	ROC_AUC SCORE
LOGISTIC REGRESSION	93.92	96.90	90.75	93.72	93.92
RANDOM FOREST	99.99	100	99.99	99.99	99.99
BAGGING CLASSIFIER	99.31	99.97	98.64	99.30	99.31
ADA BOOST	91.74	100	99.99	91.67	91.74
GRADIENT BOOST	93.48	96.65	90.07	93.25	93.48
XG BOOST	95.42	98.85	91.90	95.25	95.42

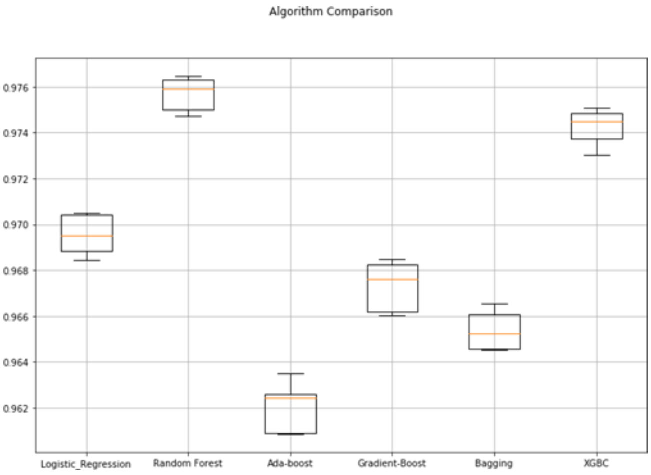




MODEL NAME(TEST)	ACCURACY	PRECISION	RECALL	FI_SCORE	ROC_AUC SCORE
LOGISTIC REGRESSION	93.99	97.08	90.71	93.78	93.99
RANDOM FOREST	95.02	99.48	90.52	94.79	95.02
BAGGING CLASSIFIER	93.45	97.69	89.01	93.15	93.45
ADA BOOST	91.76	92.60	90.77	91.68	91.76
GRADIENT BOOST	93.48	96.67	90.05	93.24	94.48
XG BOOST	95.00	98.42	91.46	94.81	95.00

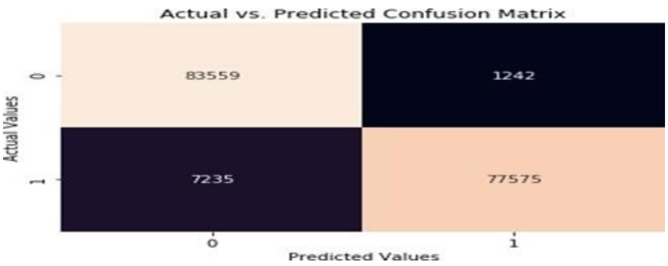


Algorithm Comparison

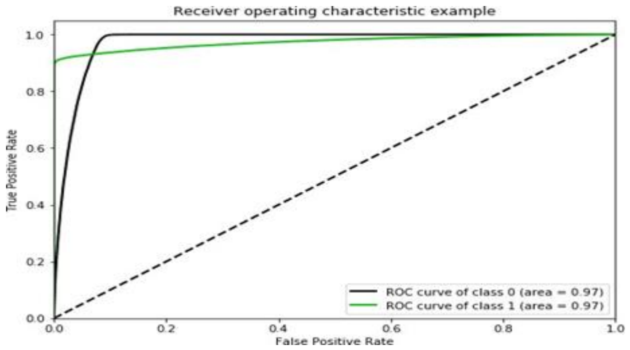


Comparisons of all models based on bias and variance errors

From the box plot comparison diagram, we can see that Random Forest and XGBC are better models amongst all of the models. But when we observe train and test scores of Random Forest and XGBC, we can see that XGBC model is performing better on both train and test data. So, we are finalizing best model as XGBC model.



Confusion Matrix of XGBC Model



Roc_Auc curve for XGBC

5. CONCLUSIONS

Credit risk assessment is merely attainable by means that of activity. Machine learning models may be used as tools to live the credit risk exposure of assorted monetary establishments. With the proper prediction of credit risk, its management can become effective and economical. This project work concentrates on the analysis varied} machine learning classifier models to predict the credit risks related to various borrowers of an establishment. For this, the main assessment parameters of the establishment are taken because the predictor variables. There are several classifier models we've approached that are mentioned within the report. we will say once and for all that XG-Boost is that the model that performed well in our project. On the opposite hand, completely different applied math techniques just like the chi-square take a look at, a pair of sample t-tests, etc. ar performed to work out the vital options. However, we've conjointly tried the K-Best technique to work out the feature importance. Feature integration has conjointly been enforced owing to its high spatial property. standardization, one-hot cryptography, and normal scalar ways are dead to visualize the advance of the model performance.

6. REFERENCES

[1] https://www.kaggle.com/c/home-credit-default-risk - Data set link.

[2] Machine Learning by Tom M Mitchell.

[3] Hands on Machine Learning with Sckit-Learn and TensorFlow by Aurelien Geron.

[4] https://scikitlearn.org/stable/modules/generated/sklearn.decomposition.PCA.htmlPCA

[5] https://scikit-learn.org/stable/supervised_learning.html#supervisedlearning-Machine Learning Models

[6] https://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.SelectK Best.html- Feature Engineering

[7] https://docs.scipy.org/doc/scipy/reference/stats.html-statistical tests

[8] https://machinelearningmastery.com/smote-oversampling-forimbalanced-classification/-smote

[9] https://machinelearningmastery.com/hyperparameters-forclassification-machine-learning-algorithms - Hyperparameter tuning

[10] https://en.wikipedia.org/wiki/Decision_tree - Decision Tree

[11] https://en.wikipedia.org/wiki/Ensemble_learning - Ensemble Learning models

[12] https://towardsdatascience.com/cross-validation-and-hyperparametertuning-how-to-optimise-your-machine-learning-model13f005af9d7d?gi=1d33882a8888 - Hyperparameter Tuning