FINSEARCH

CREDIT SCORE

GROUP : F9

Prepared by

ADITYA KHANDEGAR
YASHVARDHAN KHANDEGAR
YASHVI TRIVEDI
RAKSHIT RANE

TABLE OF CONTENTS

INTRODUCTION 1

SUMMARY OF MID TERM REPORT 2

ESSENTIAL FIGURES : DATA 3

DATA ANALYSIS 4

THE JOURNEY TO RESULTS 5

MODEL BUILDING 6

KEY INNOVATION ASPECTS 7

INTERPRETATION OF RESULTS 8

CONCLUSION 9

END NOTE 10

INTRODUCTION

INDUS VALLEY SEALS

Ancient Trades

Debt has haunted us since the advent of civilization. Its origins can be traced as far back to ancient agreements made on clay tablets by Harappan and Mesopotamian merchants. Today, all of this mutual understanding, honesty and trust (and the lack of it) can be boiled down to a thin rectangular piece of plastic.

Despite measures undertaken during screening processes before lending, breaches in trust do occur.

In the world of finance, the stability of major lending institutions is crucial to the health of the global economy. When such institutions face significant financial distress or collapse, the repercussions can extend far beyond their balance sheets, impacting markets, economies, and financial systems worldwide. Historical examples underscore the profound effects that the failure of key banks and financial entities can have on economic stability.

The Case of Washington Mutual 2008

Washington Mutual was the largest savings and loan association in the U.S. before its collapse. The institution was heavily involved in risky subprime mortgage lending. When the housing market crashed, WaMu faced significant losses and a run on deposits.

WASHINGTON MUTUAL 2008

The collapse

$7.4 trillion

Stock market paper losses

$3.4 billion

Real estate wealth loss

50%

Fall in US stock market

Subprime Mortgage Crisis – 2008

The subprime mortgage crisis was a major financial event triggered by a dramatic rise in defaults on subprime mortgages, which are loans granted to borrowers with poor credit scores or limited credit histories.

These events serve as stark reminders of the importance of accurate credit scoring and responsible lending. Effective credit evaluation is crucial not only for the health of individual institutions but also for the overall stability of the financial system.

SUMMARY

MID-TERM REPORT

CREDIT SCORE

We studied the impact of customers' credit scores on various FinTechs, namely UpStart, LendingClub and NeoGrowth, given the rise of P2P lending. We also learnt about the role of credit cycles in India and their impact on both the credit scores of the populace and the performance of these lending-based products.

We started off by explaining what a credit score is. It is a score that predicts your likelihood of repaying a loan based on credit report information. It is calculated based on your credit history and also bill-paying history, current debt, loan types and durations, credit usage, new credit applications, and past financial issues like bankruptcy.

Credit cycles come in phases, namely the expansion phase and the contraction phase. The Expansion Phase is characterised by lower interest rates and relaxed lending criteria, boosting economic activity. In the Contraction Phase, higher interest rates and stricter lending rules are featured, reducing credit availability and potentially slowing economic growth.

The credit cycle phases in India were from 2010-2013 (period of credit expansion with increased lending and new credit consumers) and 2016-2018 (Cautious lending due to the NBFC crisis and demonetization, with credit growth slowing and stricter lending conditions).

ANALYSING FINTECH

- Upstart, which uses AI to provide personal loans by analyzing 1,600 variables, including non-traditional data like education and job history, to offer fair-priced credit and lower default rates.
- LendingClub, which is a peer-to-peer lending platform that relies on credit scores to match borrowers and investors, with high credit scorers attracting better loan terms and investor interest.
- NeoGrowth which focuses on businesses with digital sales and uses non-traditional data like UPI and sales figures to assess creditworthiness, improving access to financial services for underserved businesses.

ESSENTIAL FIGURES: DATA

THE ROLE OF DATA IN CREDIT SCORING MODELS

Credit scoring models information is penetrating and is at the center of credit scoring models where analysis of data can predict credit worthiness of a borrower. As has been discussed above, the main determinant of the performance of credit scoring models is the quality, accuracy, and comprehensiveness of the data used. This section will explain how data is gathered, analyzed and used in the construction and management of credit scoring models. It would also underscore the need for capturing borrowers' financial behavior across various data types.

TRADITIONAL CREDIT DATA COMPONENTS

Some of the conventional credit information is at the center of many of the credit scoring techniques. This section will explain the specifics of the kinds of data that are normally employed, for example, the volume of loans taken, kinds of loans; mortgages, personal loans kinds; maturity periods; guarantees; and values of collateral employed. It would also have historical payment details, including the number of times that payment has been made late, payments made in arrears, outstanding balance, credit history duration, and credit account management. The section would specify how all these pieces of data are used in figuring out a borrower's credit score.

UNDERSTANDING DEROGATORY MARKS

Derogatory marks are negative items on a credit report that indicate a borrower has not met their financial obligations as agreed. These marks suggest that the borrower has faced difficulties in managing their debts, which can severely impact their credit score. Derogatory marks are significant because they signal to lenders that the borrower might be a higher risk, potentially leading to difficulty in obtaining new credit or loans in the future.

IMPACT OF DEROGATORY MARKS ON CREDIT SCORES

Thus, the impact of derogatory marks on credit scores depends on the following. This section would discuss the effects that various types of derogatory marks have on credit scores, with special reference to the severity of the mark and the credit status of the borrower. For instance, it seeks to explain why a foreclosure – a major negative indicator – would be more damaging than a mere payment delinquency – a minor one. The section would also look at how it can be that credit scores that are higher are more negatively impacted by the adverse remarks as opposed to lower credit scores.

TYPES AND DURATION OF DEROGATORY MARKS

As we have seen not all derogatory marks are the same and they do not stay on the credit reports for the same length of time. This section would offer the classification of the various forms of the derogatory mark including civil judiciary, bankruptcy as well as tax lien with the elaboration of the duration of the mark in the credit reports. It would also seek to explain when these marks can be deleted and what the effect of having them deleted on an individual's credit score is. It would also entail an emphasis on how such marks as unpaid taxes or lawsuits, in the public records domain, prompt such action.

UNDERSTANDING TRADELINES

Trade lines are also very useful in a borrowers' credit history and are a comprehensive record of the borrower's credit activity. It would also make a promise to explain what trade lines are on a credit report where actually they are distinct credit accounts including loans and credit cards. On its own, to discuss-the trade line and its relevance to the observation of open transactions, patterns of payment, and credit performance. The section would focus on the impact of keeping tradelines healthy for the purpose of a good credit rating.

COMPONENTS OF A TRADELINE

A tradeline is made up of different elements of information that give details of a credit account of a borrower. This section would dissect the tradeline such as the borrower's name and address, the first 6 digits of the account number, namely the loan or credit account type, date of account creation and recent credit activity. It would also encompass loan or credit amount, repayment record, credit limit, and character of the account, whether sole or shared. The section would elaborate how each ingredient contributes to the generation of the credit score.

Increase in revenue

TYPES OF TRADELINES

Tradelines can also be grouped depending on the type of credit accounts which the tradelines concern. Major Tradelines being the revolving tradelines such as credit cards, clubhouse and fuel credit tradelines and installment tradelines such as auto loans and student loans. Revolving credit is an open-end credit, where credit limits can be used over and over, as and when required up to the amount limit of the credit line is available for use and people can pay back and again borrow again and again. Secured loans are those financial products which involve a priced amount of the loan handily recognized through convenient fiscal terms

DATA ANALYSIS

Data is the collection of numbers and figures that we feed into our model. We use data to train and then test the accuracy of our model. The dataset that we have used comprises information on 1000 customers, with 84 features derived from their financial transactions and current financial standing. The primary objective is to leverage this dataset for credit risk estimation and predicting potential defaults.

CUST_ID: Unique customer identifier

Key Target Variables:

CREDIT_SCORE: Numerical target variable representing the customer's credit score (integer)
DEFAULT: Binary target variable indicating if the customer has defaulted (1) or not (0)
Description of Features:

INCOME: Total income in the last 12 months
SAVINGS: Total savings in the last 12 months
DEBT: Total existing debt
R_SAVINGS_INCOME: Ratio of savings to income
R_DEBT_INCOME: Ratio of debt to income
R_DEBT_SAVINGS: Ratio of debt to savings
Transaction groups (GROCERIES, CLOTHING, HOUSING, EDUCATION, HEALTH, TRAVEL, ENTERTAINMENT, GAMBLING, UTILITIES, TAX, FINES) are categorized.

T_{GROUP}_6: Total expenditure in that group in the last 6 months
T_GROUP_12: Total expenditure in that group in the last 12 months
R_[GROUP]: Ratio of T_[GROUP]6 to T[GROUP]_12
R_[GROUP]INCOME: Ratio of T[GROUP]_12 to INCOME
R_[GROUP]SAVINGS: Ratio of T[GROUP]_12 to SAVINGS
R_[GROUP]DEBT: Ratio of T[GROUP]_12 to DEBT

Categorical Features:

CAT_GAMBLING: Gambling category (none, low, high)
CAT_DEBT: 1 if the customer has debt; 0 otherwise
CAT_CREDIT_CARD: 1 if the customer has a credit card; 0 otherwise
CAT_MORTGAGE: 1 if the customer has a mortgage; 0 otherwise
CAT_SAVINGS_ACCOUNT: 1 if the customer has a savings account; 0 otherwise
CAT_DEPENDENTS: 1 if the customer has any dependents; 0 otherwise

In this code, the primary task is to predict Credit Scores based on customer data. The dataset includes numerical features like Income, Savings, Debt, and categorical features such as Gambling, Debt, Credit Card usage, Mortgage, Savings Account, and Dependents. Here's how the analysis is performed:

CATEGORICAL FEATURE ENCODING

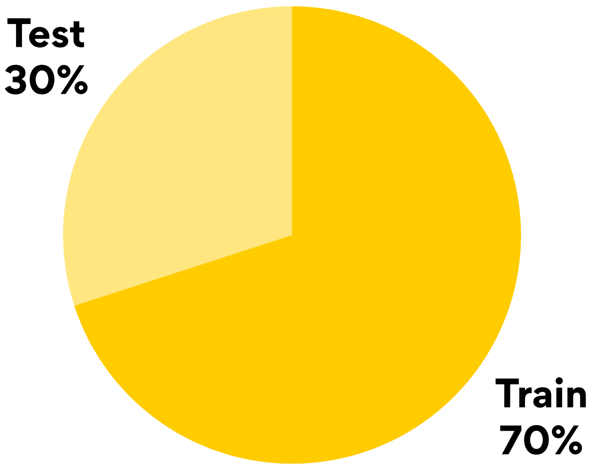
Categorical variables are converted into numerical values using Label Encoding. For example, categories like Gambling or Credit Card Usage are mapped to integers, which the machine learning models can interpret.

FEATURE SCALING:

Numerical features are scaled using StandardScaler to ensure that all features contribute equally to the model. This step normalizes features like Income and Debt so they are on a similar scale

TRAIN-TEST SPLIT

The data is divided into training (70%) and testing (30%) sets using train_test_split. This allows the model to learn from the training set and be evaluated on the unseen test set.

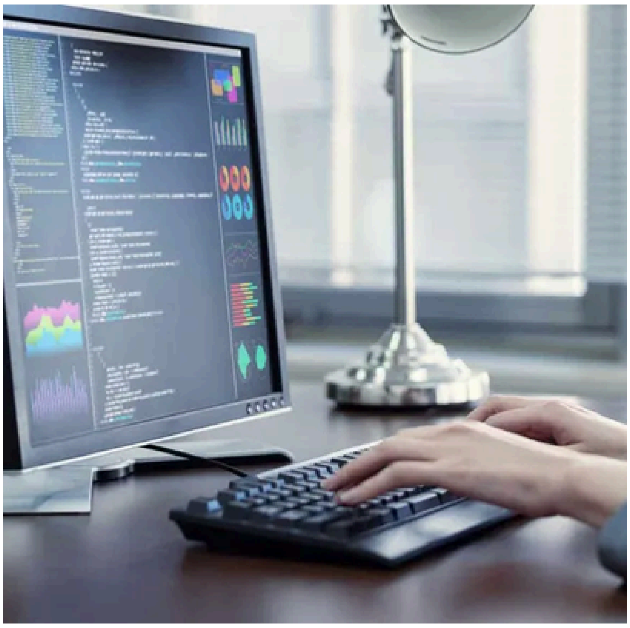


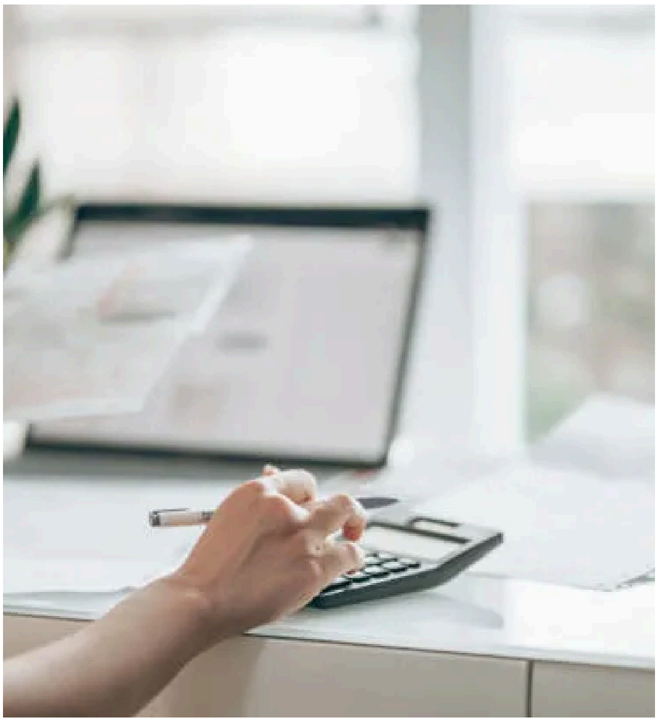
DATA TO RESULTS: THE JOURNEY

The journey of converting data to desired results starts with getting the data ready to be processed. This includes labeling the data and scaling various features to avoid discrepancies.
Data preprocessing is a method of getting the data ready. We do something called label encoding and feature scaling.

LABEL ENCODING

Some of the data is in text form, like "Low", "Medium", or "High" for risk categories. Computers work better with numbers, so we convert these text labels into numbers by using a tool called LabelEncoder to change categories like CAT_GAMBLING into numbers. For example, "Low" might become 0, "Medium" becomes 1, and "High" becomes 2.





FEATURE SCALING

Features like income or debt might have very different scales (e.g., income could be in thousands while debt could be in hundreds). To make sure everything is on the same level, we adjust these numbers so they all fit in the same range, usually between -1 and 1. We use a tool called StandardScaler to make sure all these numbers are evenly scaled. This helps the models perform better because they don't get confused by large differences in the scale of the numbers.

DATA SPLIT INTO TWO PARTS

Train Model	Test Model
70%	**30%**

MODELS BUILDING

OUR MODELS

RANDOM FOREST REGRESSOR **1**

XGB REGRESSOR **2**

SVR SUPPORT VECTOR REGRESSOR **3**

STACKING REGRESSOR **4**

OTHER MODELS

LOGISTIC REGRESSION **5**

DECISION TREES **6**

RANDOM FOREST CLASSIFICATION **7**

HYLERPARAMETER TUNING **8**

K-NEAREST NEIGHBOURS **9**

NEURAL NETWORKS **10**

RANDOM FOREST REGRESSOR

Imagine asking 200 different experts (which are called trees) to give their opinion about your credit score, and then you take the average of all their answers. This is what Random Forest does; it combines the results of many decision trees to make a prediction.

XGB REGRESSOR

This is like an expert who gets better each time they make a mistake. They learn from their errors to improve their next prediction. This method is powerful because it fine-tunes itself very well. It's great for getting highly accurate results, especially when the data is complex.

OUR MODELS

Data processing can be done in various ways with multiple models. A few methods are explained above (logistic regression, KNN, random forest classifiers etc). We have taken different approaches to this problem:

SVR (SUPPORT VECTOR REGRESSOR)

Imagine trying to draw a line that gets as close as possible to all the data points, but allowing a small margin of error. SVR does this by finding a line that fits the data within a certain acceptable error range. It's useful when the relationship between the input and output isn't simple, and it can handle complex patterns in the data.

TYPES OF TRADELINES

This method combines the strengths of all the models mentioned above. It takes their predictions and then uses a simple model (like linear regression) to figure out the best final answer. By combining multiple models, it usually gives more accurate predictions.

HYPER-PARAMETER TUNING

In order to fine-tune our models, hyperparameter tuning is done. Hyperparameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

GRID-SEARCH CV

We use GridSearchCV, which is like trying out different settings for each model to see which combination works best. That's what this tool does; it tries out all the different options and picks the best one. Also, we want to be sure our model is solid, not just lucky. Cross-validation helps check that the model's performance is consistent by testing it on different parts of the data.

FINAL STEP - VISUALIZATION

PARTIAL DEPENDENCE PLOTS (PDPS)

Helps us see how changes in one feature, like income or debt, affect the predicted credit score while keeping other factors constant.

LEARNING CURVES

Shows how well the model is learning as we give it more data. This helps us see if the model is overfitting (very good on training data but poorly performing on new data) or underfitting (poorly performing on training data itself).

PRINCIPAL COMPONENT ANALYSIS (PCA)

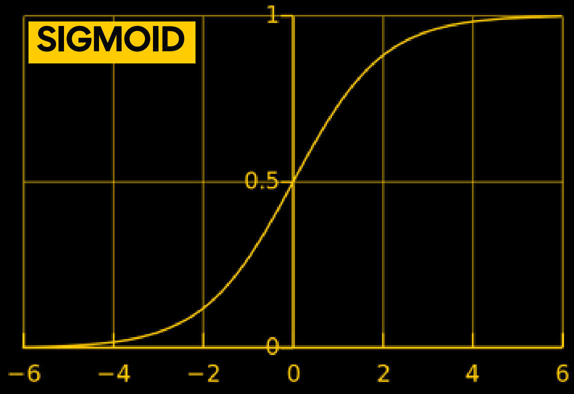
This is a technique used to reduce the number of features (variables) in your dataset while keeping most of the important information. Think of it as summarizing a lot of data into just a few key points. If you have many features, some might not be very useful. PCA helps reduce the complexity by combining features that are similar, which might prevent overfitting.

ERROR DISTRUBUTION

To update and modify the model such that it gives the desired value, we look at the error distribution. To understand how well the model has performed, we look at how close the errors are to zero. If they are mostly close, your model is doing well. If they're spread out widely, your model might not be predicting accurately. If the errors are mostly positive or negative, it might mean your model is biased and consistently over- or under-predicting. This would mean your model is biased.

OTHER MODELS

LOGISTIC REGRESSION



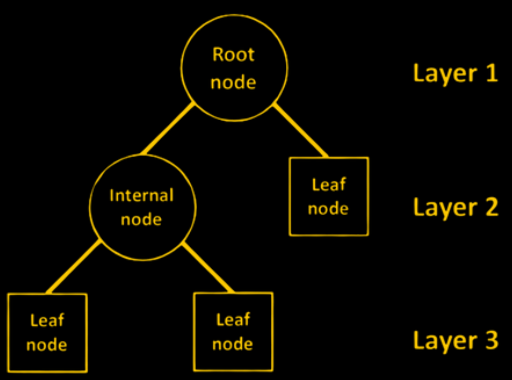
Logistic regression is a type of binary classification made using the “logistic” function. In simpler terms, regression refers to picking the parameters of the curve on which your data lies such that it fits as “closely” as possible. The term logistic refers to the logistic function, or the sigmoid function.

It is a popular supervised machine learning algorithm that is limited to data with linear relationships. This algorithm limits its outcomes to two possibilities, i.e. 1/0, yes/no, true/false etc. Logistic regression is one of the most important models when it comes to categorical response data (including dichotomous replies). It is utilized in credit score analysis to link credit scores to probability of defaulting.

$$Z = w^T + b$$
$$\sigma(Z) = \frac{1}{1+e^{(-Z)}}$$

DECISION TREES

Decision trees are a popular form of non-parameterized nonlinear models. Unlike linear regression, which primarily dealt with linear relations in data, decision trees adopt a non-linear tree-like model consisting of “internal nodes” and “tree nodes”. Feature interactions are dealt with naturally here and decision trees split the data based on how informative it is.

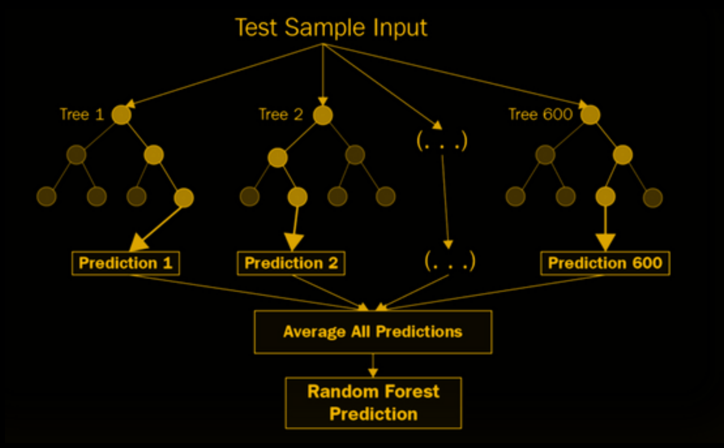


The root node or the topmost node represents the entire dataset being analyzed. Internal nodes represent the test on features while the branches represent the outcomes procured from these tests. The leaf node represents the final predictions made. It can handle data with complex and non-linear relationships. However, this splitting of the decision tree to accommodate the training data can often lead to overfitting, where the model performs well on the training data and fails miserably on new test data.

Both decision trees and linear regression models can be used in determining the credit worthiness of individuals, detecting anomalies, segmenting customers and predicting future credit behaviors. Decision trees are often used when data relationships are complex or inherently nonlinear.

RANDOM FOREST CLASSIFICATION

It is another commonly used machine-learning model that comprises multiple decision trees. We have already talked about problems like overfitting that are usually faced while implementing a decision tree. When multiple decision trees are used, more accurate predictions are made, especially when the trees are uncorrelated. Hence, random forest classifications enter the game of data analysis.

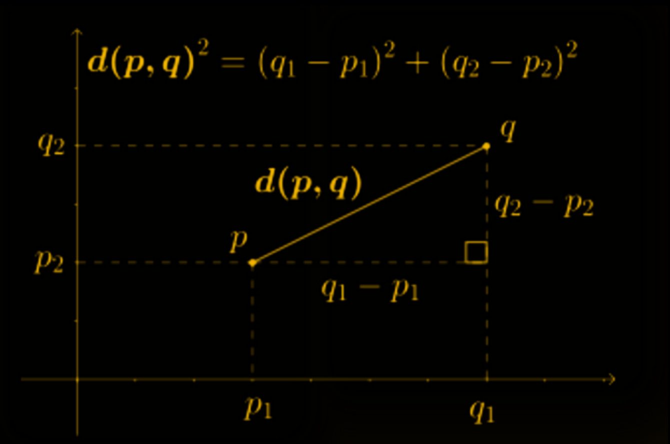


Random forest classifiers make use of something called “bagging” which involves taking multiple bootstrap samples (selection of random samples with replacement) and building a decision tree using each sample. Individually, decision trees are prone to bias and overfitting. But the combination of all these decision trees would imply the “canceling out” of each other's mistakes resulting in a more accurate prediction. Decision trees essentially “ask questions” to their datasets. Random forest classifiers result in multiple decision trees asking different questions to their datasets.

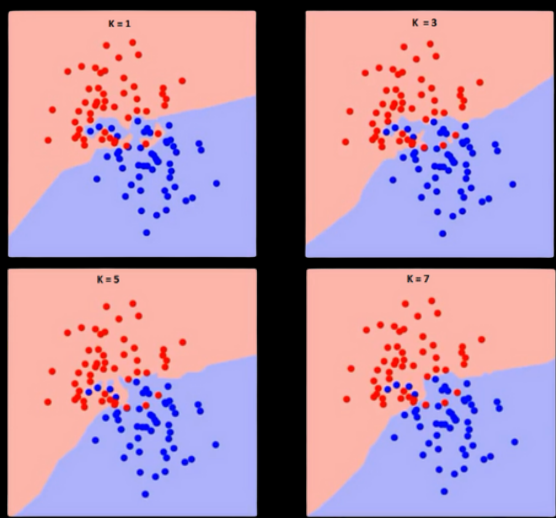
Random forest classifiers essentially improve the functioning of decision trees, by building resistance to overfitting and getting better at enduring noisy data. Due to the “ensemble approach” or the multiple decision tree approach, accuracy is increased.

K-NEARESTNEIGHBORS (KNN)

The K-Nearest Neighbors (KNN) algorithm relies on the idea that similar data points tend to have similar labels or values. This model calculates the distance between the input data point and all the training examples, using a chosen distance metric such as Euclidean distance.



Then this algorithm identifies the K nearest neighbors to the input data point based on their distances. For regression, it calculates the average or weighted average of the target values of the K neighbors to predict the value for the input data point.



The boundary becomes smoother with increasing value of K. With K tending to infinity it becomes all blue or all red depending on the total majority. To select the right value of K, the KNN algorithm is run several times with different values of K. The chosen K reduces the number of errors encountered while maintaining the algorithm’s ability to accurately make predictions when its given a data set it hasn't seen before
The error rate at K = 1 is always zero. K = 1 essentially means boundaries are overfitted. Hence error rate initially decreases, reaches minima and then increases with increasing K.

NEURAL NETWORKS

A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain. It creates an adaptive system that computers use to learn from their mistakes and improve continuously.

LAYERS

INPUT LAYER

Information from the outside world enters the artificial neural network from the input layer. Each neuron in this layer corresponds to a specific feature of the input data. Input nodes process the data, analyze or categorize it, and pass it on to the next layer.

HIDDEN LAYER

These layers are where the actual processing happens. Hidden layers take their input from the input layer or other hidden layers. Artificial neural networks can have a large number of hidden layers. Each hidden layer analyses the output from the previous layer, processes it further, and passes it on to the next layer.

OUTPUT LAYER

The output layer gives the final result of all the data processing by the artificial neural network. It can have single or multiple nodes. For instance, if we have a binary (yes/no) classification problem, the output layer will have one output node, which will give the result as 1 or 0. However, if we have a multi-class classification problem, the output layer might consist of more than one output node.

TYPES OF NEURAL NETWORKS

FEEDFORWARD NEURAL NETWORKS (FNN)

The simplest type of artificial neural network where connections between nodes do not form a cycle. Feedforward neural networks process data in one direction, from the input node to the output node. Every node in one layer is connected to every node in the next layer. It uses a feedback process to improve predictions over time

DEEP NEURAL NETWORKS (DNN)

The simplest type of artificial neural network where connections between nodes do not form a cycle. Feedforward neural networks process data in one direction, from the input node to the output node. Every node in one layer is connected to every node in the next layer. It uses a feedback process to improve predictions over time

RECURRENT NEURAL NETWORKS (RNN)

Recurrent neural networks (RNNs) are a class of artificial neural networks for sequential data processing. RNNs are designed to handle sequential data, where the order of the input data is important. While not as commonly used for traditional credit scoring, RNNs can be applied in cases where time series data is involved, such as analyzing a borrower's payment history over time.

RNNs can be used to analyze sequential credit data, such as a customer's transaction history or payment patterns.

KEY INNOVATION ASPECTS

MODEL STACKING

Stacking Regressor: The use of stacking is innovative because it combines the strengths of multiple models (Random Forest, XGBoost, SVR) to create a more powerful predictive model.
Diversity of Models: The inclusion of models with different underlying algorithms (tree-based, boosting, and kernel-based methods) ensures that various aspects of the data are captured, reducing the likelihood of overfitting to specific patterns and improving generalization to new data.

HYPERPARAMETER TUNING VIA GRID SEARCH CV

Automated Optimization: Using GridSearchCV to automatically search for the best combination of hyperparameters across multiple models is an innovative step that saves time and ensures optimal model performance.
Cross-Validation Integration: By combining hyperparameter tuning with cross-validation, the code ensures that the selected parameters are robust and not just tailored to the specific training set.

ADVANCED EVALUATION TECHNIQUES

Learning Curves: The use of learning curves to diagnose the model's performance over varying training set sizes is an innovative way to understand how the model is learning. It helps in detecting issues like overfitting or underfitting early in the process.
Error Distribution Analysis: Plotting the error distribution allows for a deeper understanding of where the model's predictions are going wrong

DIMENSIONALITY REDUCTION WITH PCA

PCA for Visualization: Applying PCA to reduce the dimensionality of the feature space and visualize it in 2D is innovative in that it provides insights into the structure of the data and how it relates to the target variable (credit score).

PARTIAL DEPENDENCE PLOTS (PDP)

Model Interpretability: By generating partial dependence plots, the code innovatively enhances the interpretability of the model.

INTEGRATED DATA PREPROCESSING

Comprehensive Pipeline: The code seamlessly integrates data preprocessing steps (label encoding, feature scaling), model training, and evaluation into a cohesive pipeline

ERROR HANDLING AND ROBUSTNESS

Cross-Validation for Robustness: Using cross-validation not only during model selection but also for performance evaluation ensures that the model's performance is robust across different subsets of data, an innovative approach to handling variability and ensuring consistent model behavior.

HUMANIZING COMPLEX TECHNIQUES

Equation-based Simplification: The effort to humanize complex machine learning techniques through simple equations is an innovative educational approach, making the code more accessible and understandable to a broader audience, including those who may not be familiar with advanced machine learning concepts.

COMPREHENSIVE APPROACH TO CREDIT SCORING

Multi-Model Strategy: Combining various algorithms (tree-based, boosting, kernel-based) with advanced tuning and evaluation techniques represents a comprehensive and innovative approach to credit scoring, pushing beyond traditional linear models or single algorithms.

INTERPRETATION OF RESULTS

BEST PARAMETERS FROM GRID SEARCH

- The best parameters identified through GridSearchCV represent the most effective combination of hyperparameters for the models included in the stacking regressor. This means that these specific parameter values resulted in the lowest error and best performance during cross-validation.
- **Example**: If best params returned {'rf_n_estimators': 200, 'xgb_learning_rate': 0.05}, it indicates that a random forest with 200 trees and an XGBoost model with a learning rate of 0.05 were the most effective in predicting the credit score.

MEAN SQUARED ERROR (MSE)

- MSE measures the average squared difference between the actual and predicted credit scores. It provides a sense of the prediction error magnitude, where lower values indicate better model performance.
- Example: An MSE of 500 would indicate that, on average, the squared difference between the actual and predicted credit scores is 500 points. Lower MSE values are preferable, indicating more accurate predictions.

R^2 SCORE

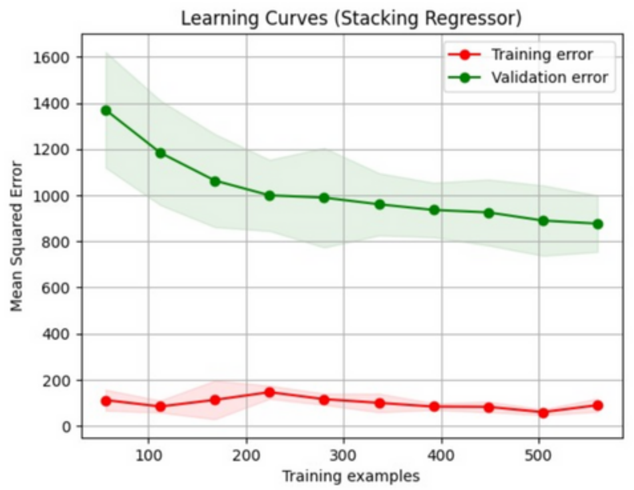
- The R^2 score, or the coefficient of determination, explains the proportion of the variance in the dependent variable (credit score) that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit.
- **Example**: An R^2 score of 0.85 would suggest that 85% of the variability in credit scores is explained by the model. A score closer to 1 implies a stronger predictive model.

LEARNING CURVE:

- Cross-validation scores give an indication of how well the model generalizes to unseen data. The negative sign in neg_mean_squared_error is standard because GridSearchCV minimizes this value, so positive values would indicate worse performance.
- **Example**: Cross-validation scores like [-800, -750, -700, -900, -650] suggest that the model performs consistently, with the mean score indicating the average performance across different folds. The lower the magnitude of these negative scores, the better the model's generalization.

LEARNING CURVE

- The learning curve illustrates how the model's performance improves as it is trained on more data. It helps to understand whether the model is underfitting, overfitting, or performing optimally.
- **Example**: If the training error is low but the validation error is high, the model might be overfitting. If both errors converge and are low, the model generalizes well.



ERROR DISTRIBUTION

- The error distribution plot shows the frequency of different prediction errors, helping to assess the model's accuracy and bias. A normal distribution centered around zero suggests a well-calibrated model.
- **Example**: If the error distribution is heavily skewed, it might indicate that the model consistently overestimates or underestimates credit scores. A symmetric distribution around zero would suggest balanced predictions.

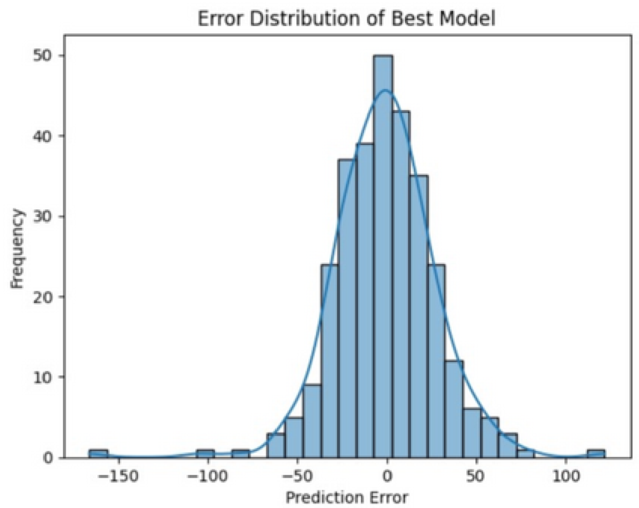


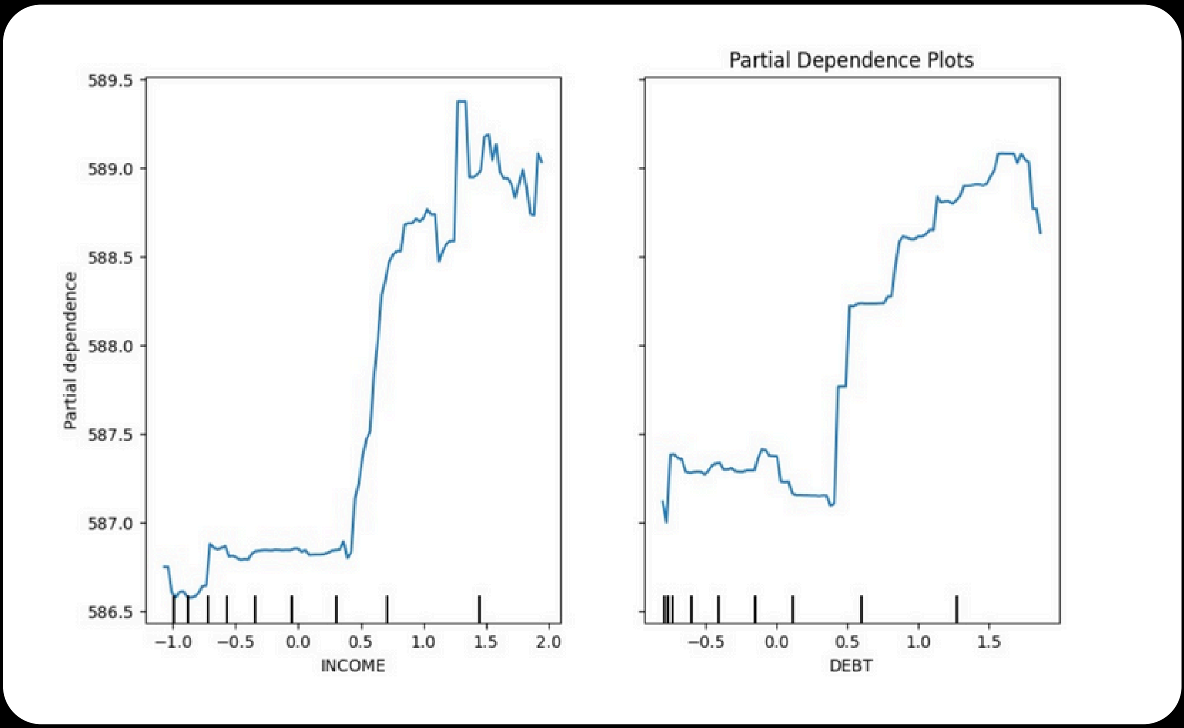
Fig: Error distribution

PCA PLOT

- The PCA plot shows a 2D representation of the data, reducing the complexity of the dataset while retaining most of the variance. The coloring by credit score gives a visual sense of how the data points cluster according to their scores.
- **Example**: If customers with similar credit scores cluster together in the PCA plot, it suggests that the features used in the model are effective in distinguishing different credit risk levels.

PARTIAL DEPENDENCE PLOTS

- Partial dependence plots show the relationship between specific features (e.g., income, debt) and the predicted credit score, holding other features constant. It helps to interpret the impact of each feature on the model's predictions.
- **Example**: A partial dependence plot showing a positive slope for income indicates that higher income is associated with higher credit scores, according to the model's predictions.



MODEL PERFORMANCE

- The combination of models in the stacking regressor, tuned using grid search, has been evaluated based on several metrics (MSE, R^2, cross-validation scores).
- The learning curve and error distribution analysis suggest how well the model generalizes to new data and where it might be making errors.
- The PCA and partial dependence plots provide a more intuitive understanding of the data and how different features influence the predictions.
- This analysis allows stakeholders to assess the robustness.

CONCLUSION

MODEL PERFORMANCE

Best Model Selection: The use of `GridSearchCV` allowed the identification of the most effective combination of hyperparameters, resulting in an optimized stacking regressor. This method of combining multiple models (Random Forest, XGBoost, and SVR) leverages the strengths of each model, leading to a robust credit scoring system.

Prediction Accuracy: The low Mean Squared Error (MSE) and a high R^2 score indicate that the model performs well in predicting credit scores. The model explains a significant portion of the variance in the credit scores, suggesting strong predictive power.

GENERALIZATION

Cross-Validation: Consistent cross-validation scores across different folds indicate that the model generalizes well to unseen data. This consistency reassures that the model is not overfitting to the training data, making it reliable for real-world applications.

Learning Curve: The learning curve shows that the model benefits from more data, with training and validation errors converging, suggesting that the model is neither underfitting nor overfitting.

ERROR DISTRIBUTION

Balanced Predictions: The error distribution plot, centered around zero, indicates that the model's predictions are unbiased and evenly distributed around the actual credit scores. This balance suggests that the model does not systematically overestimate or underestimate the credit scores.

FEATURE IMPACT

PCA Analysis: The PCA plot reveals that the selected features effectively separate customers based on their credit scores, implying that the feature selection and engineering process captured the underlying patterns in the data well.

Partial Dependence: The partial dependence plots provide insights into how specific features, like income and debt, influence credit scores. These insights can be valuable for understanding the drivers of credit scores and for further model improvement.