EXPERIMENT REPORT

Student Name	Monali Patil
Project Name	AT1 - Kaggle Competition
Date	25/08/2023
Deliverables	Notebook name: Patil_Monali-14370946-week2_EDA.ipynb Patil_Monali-14370946-week2_RF.ipynb Model name: Random Forest Classifier GitHub Repo: https://github.com/MonaliPatil19/adv_mla_assignment1

---

1. EXPERIMENT BACKGROUND

Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach.

1.a. Business Objective

Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results?

Goal: The project involves building a binary classifier model that predicts whether a college basketball player will be drafted into the NBA league based on their statistics records. The dataset provided contains a wide range of features that illuminate players' performance during their college basketball season comprising 64 players' performance related attributes.

Application: This event captures the attention of sports commentators, fans, and enthusiasts who eagerly track the careers of college players and speculate about their chances of being selected by NBA teams. So, the model results will be used to assist sports commentators, fans, and teams in understanding the likelihood of a player's selection.

Impact: The model's accurate predictions can provide valuable insights for both players and teams, aiding decision-making during the NBA draft process. Additionally, this model will offer valuable insights to sports commentators, fans, and scouts, aiding them in predicting individual players' potential NBA draft prospects. Incorrect predictions, on the other hand, might lead to missed opportunities for players and teams, potentially impacting the overall performance and composition of the NBA teams.

1.b. Hypothesis	Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it, Hypothesis: To test whether the statistical attributes of college basketball players from their college season can effectively predict whether they will be drafted into the NBA league. The key question would be, can a machine learning model accurately predict whether a college basketball player will be selected in the NBA draft based on their performance metrics? In this experiment, employing random forest classifier, the aim is to assess the predictive power of these player statistics and evaluate whether they hold valuable insights into a player's likelihood of getting drafted. Rationale: Considering this hypothesis is worthwhile because the NBA draft is a significant event in a basketball player's career, and accurately predicting draft selection can have multiple benefits. For teams, it can provide a data-driven approach to identify potential talents and make informed decisions during the draft process. For players, it offers insights into the factors that influence their selection, potentially guiding their training and career decisions. Additionally, for fans and commentators, accurate predictions can enhance engagement and discussions around player performances and draft prospects. The use of statistical attributes also aligns with the growing trend of data-driven decision-making in sports, making this experiment relevant and valuable.
1.c. Experiment Objective	Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment. The expected outcome of this experiment would be the development of a binary classifier model utilizing Random Forest Classifier algorithm that can accurately predict whether a college basketball player will be drafted to the NBA league based on their college season statistics. The model would be trained on a dataset of college players information and their performance statistics, and it would use Random Forest Classification algorithm to identify patterns and make accurate predictions. The goal of this experiment would be to identify potential player for an annual event aimed at identifying the players to be drafted in the NBA league utilising Random Forest Classifier algorithm. The potential scenarios arising from this experiment include the following situations: • The model accurately predicts players who will be drafted. This provides teams with valuable insights for selecting promising talents, potentially improving their performance and team success. • The model inaccurately predicts players who will be drafted. Teams may miss out on potential talents, impacting their competitiveness. False positives rate can lead to misallocation of resources. • If the model performs moderately, there's an opportunity to fine-tune it. However, the model's marginal performance would be similar to random guessing. • The model is not accurate enough to be useful for predicting players position, and the experiment is unsuccessful. In this case, the NBA team may need to explore other methods for identifying potential players for their annual event.

2. EXPERIMENT DETAILS

Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them.

2.a. Data Preparation

Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments

The players' statistical data was explored, cleaned, and prepared through the subsequent activities for the binary classification algorithm to be employed.

• Data Understanding

1] Loading Data: Imported data from the given CSV file into the panda's data frames to use and create a binary classification model.

2] Exploring Data: Examined and studied players' information using different pandas functions to comprehend and uncover patterns, aiming to identify prospective players for NBA draft selection.

* Additionally, to ensure the quality of the data to be utilized by the model analysis, conducted below activities.

- Handing missing/null values.
- Eliminating identifiers.
- Duplicate records.
- Analysing the correlation between predictors and the target variable.
- Data distribution of various features.
- Selecting appropriate features based on the correlation coefficient.
- Accessing if imbalance targets classes.

df.head(): Checking initial records of the dataset.
df.shape(): Verifying the dimension of the dataset.
df.columns: Identifying attributes name.
df.info(): Assessing the summary information of the attributes of the dataset.
df.describe(include='all'): Examining statistical summary information for all variables of the dataset across different data types.
df.isnull().sum(): Examining whether there are any null values in the dataset.
df.duplicated().sum(): Identifying whether there are any duplicate values in the dataset.

• Data Preparation

3] Handling Missing Values

A significant number of 33 attributes in the training dataset and 24 in the testing dataset possess missing values, necessitating addressing to ensure data readiness for modelling as machine learning algorithms are incapable of processing features with missing values.

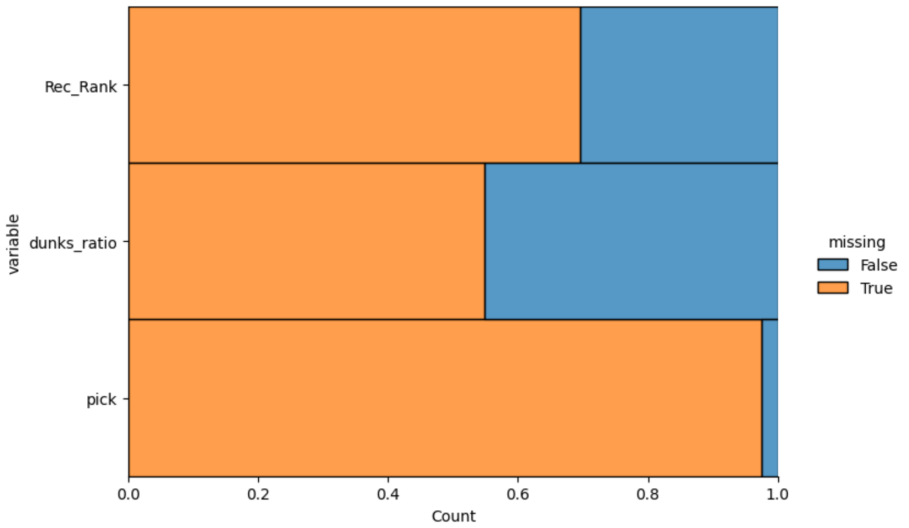

Figure 1: Attributes with majority of missing values.

* Among the features with missing data, the following three attributes exhibit notably high levels of missing values, exceeding 50%.

- Rec_Rank: Training -> 69.55% (39055/56091) and Testing -> 71.15% (3536/4970)
- dunks_ratio: Training -> 54.89% (30793/56091) and Testing -> 54.64% (2717/4970)
- pick: Training -> 97.64% (54705/56091) and Testing -> 98.98% (4921/4970)

* Therefore, it is practical to exclude these features in order to prevent potential biases in the model arising from imputation.

* Given that the remaining features have missing values comprising less than 2%, it is a reasonable approach to fill these missing values using the mean for numerical attributes and the mode for categorical attributes.

4] Verifying and Removing Identifiers

* Attributes 'player_id' and 'num' are unique identifiers for each basketball player, and its inclusion in the analysis can lead to overfitting, where the model fits to these specific values rather than the underlying generalized patterns in the sportsman's records.

* Additionally, 'team' and 'conf' attributes representing the names of teams and conferences, are overly specific and do not contribute to the model's general learning.

* Therefore, removed 'player_id', 'num', 'team' and 'conf' features from the datasets.

5] Analysing correlation between predictors and target variable

* Performing correlation analysis to determine which features would be most valuable for building a predictive model for the NBA draft prediction task.

* Correlation using heatmap, calculates the correlation between each feature and the target variable 'drafted' to understand how much they influence the target. Features with high correlations to the target variable are potentially valuable and are selected as predictors.

6] Accessing if imbalance targets classes

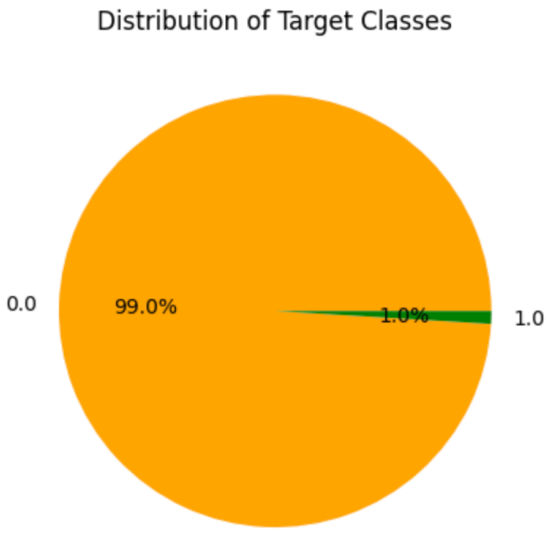

Figure 2: Target classes distribution.

* The pie chart above illustrates a significant class imbalance within the dataset. The majority of observations are attributed to a single target class, representing players who have not been drafted denoted by the value 0.

* Performed Oversampling with SMOTE (Synthetic Minority Over-sampling Technique) method to address class imbalance by generating synthetic observations for the minority class.

* It works by creating synthetic samples that are similar to existing observations in the minority class without duplicating the existing observations, rather it generates new observations by interpolating between existing ones.

7] Features Scaling

* Using feature scaling prevents the algorithm from prioritizing high-value features over other more informative ones. It ensures uniformity in feature values, enabling the algorithm to learn generalized patterns from all features for accurate player identification and predictions.

* Employed StandardScaler method because it maintains the features data distribution's shape and retains outliers by scaling data using the mean of 0 and standard deviation of 1 across the entire dataset, rather than for individual data points.

8] Hyperparameters Tuning

* Employed Hyperparameter Tuning to determine optimal values for the hyperparameters of the Random Forest Classifier algorithm to further address slight overfitting nature of the mode.

* Utilizing the Random Search technique for hyperparameter tuning, which introduces randomness by selecting values from the search space. This can lead to combinations of hyperparameter values that minimize errors.

Since the 'ht' feature, which represents player height, contains a below number of unique values, extracting accurate and up-to-date information from it is uncertain. So have excluded it.
['2-Jun','4-Jun','8-Jun','1-Jun','5-Jun','Jun-00','6-Jun','9-Jun','3-Jun','11-Jun',
'7-Jun','10-May','10-Jun','11-May','9-May','Jul-00','7-May','5-Jul','8-May',
'6-May','2-Jul','1-Jul','-','3-May','3-Jul','Apr-00','5-May','4-Jul',nan,'So', 'Jr',
'Fr',"6'4",'None','4-May','0','1-May','6-Jul','5-Apr', '2-May']

The below are essential measures that could hold significance for any future classification experiments.

1. In order to avoid the model from overfitting on specific data points and to facilitate its learning of generalized patterns, it is crucial to remove any identification attributes.
2. Thoroughly assessing and managing missing values is essential to prevent introducing biases into the model.
3. Choosing the appropriate feature scaling method among MinMaxScaler, MaxAbsScaler, or StandardScaler should depend on the presence of natural data ordering and considerations like outlier preservation.
4. In classification tasks, particularly when dealing with imbalanced data, it's crucial to maintain the same target variable's class frequencies in each set accurately.

2.b. Feature Engineering

Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments

Feature engineering was not performed and kindly refer to 4] Verifying and Removing Identifiers from above section 2.a for the rationale behind eliminating certain features.

2.c. Modelling

Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments

The target variable is binary, containing 0 signifying players who are not selected, and 1 representing drafted players. With our goal to predict outcomes within these two classes, a Binary Classification model is appropriate, aligning with the objective of categorizing players into these distinct groups.

In this experiment, utilised the Random Forest Classifier algorithm, which is well-suited for binary classification, to accurately determine whether the player will be selected for the NBA league.

Hyperparameters Selected.

Hyperparameters Name	Values Tested
class_weight:	balanced
criterion:	entropy
max_features:	log2
max_depth:	487
min_samples_leaf:	14
n_estimators:	209

	* Began with the default hyperparameters of the Random Forest algorithm and assessed its performance. The above provided hyperparameter values were employed to address the minor overfitting observed in the model. * Setting class_weight hyperparameter to 'balanced' adjusts the weights of classes in the model, which is crucial when dealing with imbalanced datasets. * Utilized 'entropy' as the criterion, for the model to prioritizes attribute selection that maximizes information gain, leading to better decision tree splits and improved overall performance. * The max_depth hyperparameter allows the model to build deeper trees, potentially capturing more complex relationships in the data and enhancing predictive power. * Selecting 'log2' for max_features hyperparameter limits the number of features considered for each split, which helps prevent the model from becoming overly specialized to noisy features. * The min_samples_leaf hyperparameter prevents overfitting by requiring a minimum number of samples in a leaf node, promoting smoother decision boundaries. * The number of estimators (trees) with n_estimators improves the model's ability to generalize by combining predictions from multiple trees. Collectively, these tuned hyperparameter values optimize the Random Forest classification model's performance by addressing issues like overfitting. Therefore, utilized Hyperparameter Tuning to identify the best hyperparameter values for the Random Forest Classifier algorithm, addressing the slight overfitting observed in the model. Random Search: Applied the Random Search method for hyperparameter tuning, which involves incorporating randomness by selecting values from the search space. This approach can generate various combinations of hyperparameter values that lead to achieving the minimal error. Grid Search's evenly spaced points may miss optimal hyperparameters. Feature Importance: Derived feature importance measure that can offer valuable insights into the players records, to help comprehend which features have the most significant impact on the model's predictions to identify prospective players for selection in the NBA league. Due to time constraints, there wasn't sufficient opportunity to thoroughly extract accurate and up-to-date information from the ‘ht’ predictor which indicates height of the players.

3. EXPERIMENT RESULTS

Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified.

3.a. Technical Performance

Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.

Baseline AUROC Performance: 0.5

* From the 1st experiment, the AUROC performance score on the validation set 0.9912 is slightly lower compared to the ideal 1.0 score achieved on the training set, indicating that the model is narrowly overfitting.

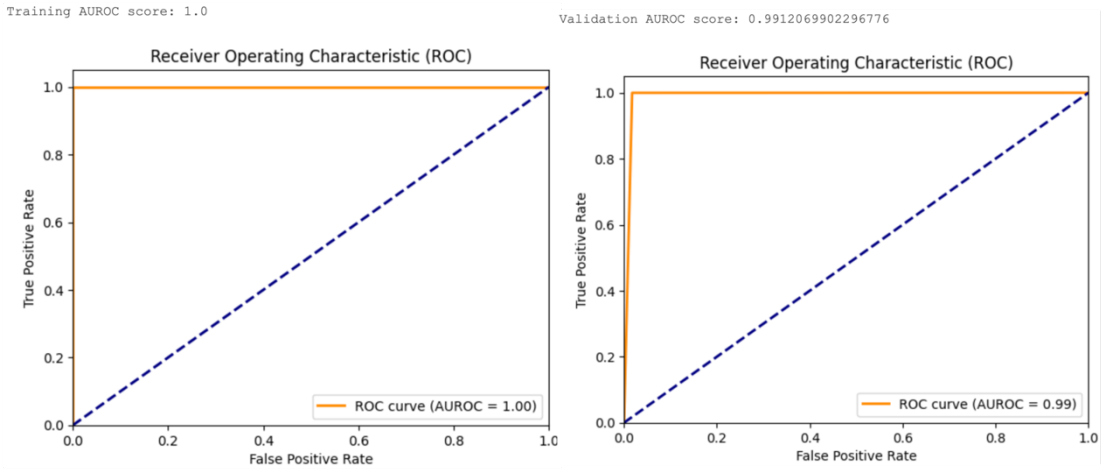

Figure 3: ROC plot from the 1st experiment.

* Additionally, from the ROC curve of training and validation datasets, visually representing the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) for different threshold values, suggest the similar light overfitting nature of the model.

* Therefore, in the 2nd experiment, Hyperparameter Tuning was applied to identify the best hyperparameter values for the Random Forest Classifier algorithm, aiming to mitigate the minor overfitting observed in the model.

Below are the Hyperparameters tuned values of the Random Forest classifier:
{'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': 487, 'max_features': 'log2', 'min_samples_leaf': 14, 'n_estimators': 209}

* Utilising Hyper tuned parameters in the 2nd experiment, there is slight variation in AUROC scores between training (0.9879) and validation (0.9844) datasets suggests that the model with optimized hyperparameters maintains consistent performance across datasets with minor differences.

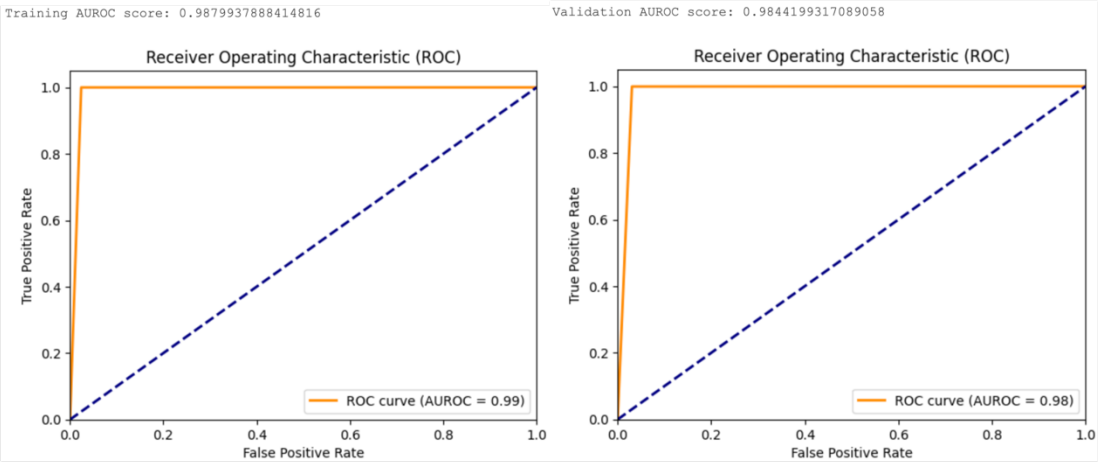

Figure 4: ROC plot from the 2nd experiment.

* The ROC plot for both training and validation datasets is relatively closer to curve of the upper left corner, illustrating that the model is behaving almost uniformly on both datasets with minute variation.

* However, there is a minor decrease in the performance metric compared the model from 1st experiment, signalling instances where players were misclassified as drafted, affecting the NAB team's overall performance and the model is not capturing important patterns and not generalised enough to identify potential players that will be drafted in the NBA league.

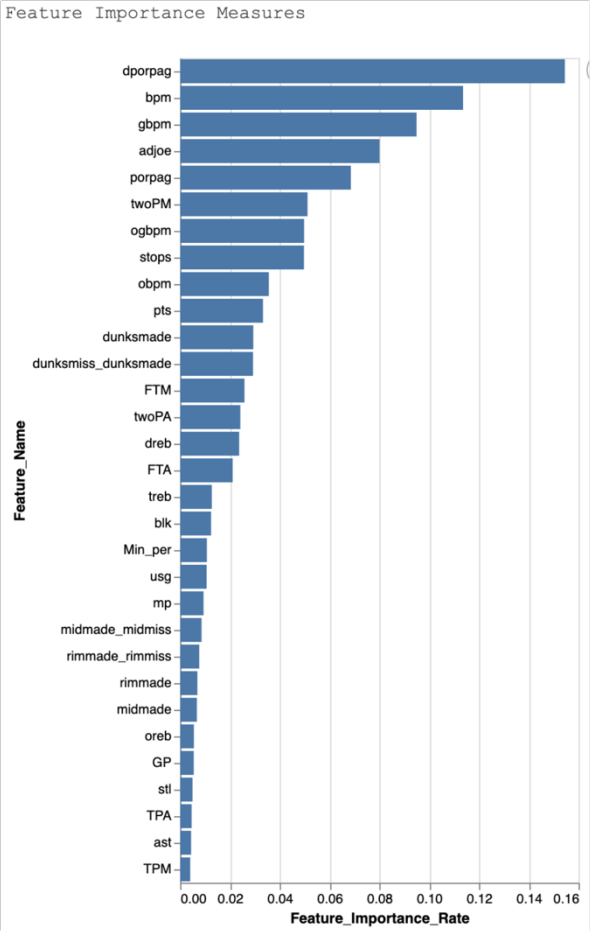

Figure 5: Feature Importance Measures.

* The bar chart above illustrates that the features making the least contribution include 'TPM',

	'ast', 'TPA', 'stl', 'GP', 'oreb', 'midmade', 'rimmade', 'rimmade_rimmiss', 'midmade_midmiss', 'mp', 'usg', 'Min_per', 'blk', 'treb'. * And the rest of the features have a significant impact on predicting either class 1, indicating players selected for the NBA league, or class 0, representing those not chosen. Their feature importance rates range from 0.02% to 0.15%.
3.b. Business Impact	Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others) The AUROC score of 0.9879 for training and 0.9744 for the validation dataset indicates a consistent prediction of drafted players across both datasets. However, the model's performance is relatively degraded, suggesting that it lacks the accuracy required to be effective in predicting player positions during the annual draft event. Additionally, the model could be further fine-tuned, optimizing its performance by leveraging feature importance assessment measures. This would ensure that the predictive classification model offers valuable insights to NBA teams, aiding in the selection of promising talents and potentially enhancing both team performance and overall success.
3.c. Encountered Issues	List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments. Considering the context of problem and the business view, height of the players from the 'ht' feature is one of the significant attribute determining their selection in the NBA league. However, it contains number of unique values and extracting accurate and up-to-date information from it is unclear. However, upon through analysis of the feature values, tried multiple logic such as months of the year would be height in feet and day of the month to be inches, but due to time constraints, there wasn't sufficient opportunity to thoroughly extract accurate information.

---

4. FUTURE EXPERIMENT	
Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective.	
4.a. Key Learning	Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end. Utilizing hyper-tuned parameters in the 2nd experiment, the model exhibited consistent behaviour across both datasets. However, its performance experienced a minor decline in comparison to 1st experiment, achieving an AUROC score of 0.9879 for training and 0.9844 for validation indicating that the model's ability to generalize and accurately predict potential drafted

	players on new, unseen data may be slightly compromised. Additionally, evaluating both the performance metric and the ROC curve plot indicates that the Logistic Regression Classifier model from the first week's experiment outperforms and is more suitable compared to the Random Forest Classifier model used in the second week.
4.b. Suggestions / Recommendations	Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production. Before considering the deployment of the model in the operational environment, it's advisable to evaluate the performance of alternative algorithms using the Feature Importance information and then select the most effective model. Thus, in the next experiment, I would utilize the Feature Importance information with different Classifier model to verify if the model's generalized performance could be enhanced. Moreover, I intend to address and extract valuable insights from the 'ht' feature, aiming to enhance the model's overall performance.