A STUDY INTO PERFORMANCE OF DIFFERENT MACHINE LEARNING METHODS IN THREE DATASETS- CHESS GAME, USED CARS AND A JOB CHANGE OF DATASCIENTIST

Iswarya Yogeashwaran
X20155034
MSc. Data Analytics
National College of Ireland
Data Mining and Machine Learning-1

***Abstract*— In this project, three different datasets such as chess game, used cars and a job change of Data Scientists are chosen to perform machine learning algorithms. The machine learning techniques used in the three datasets are Random forest, Decision tree, Xgbboost (eXtreme Gradient Boosting), and KNN (K-Nearest Neighbours). This project explains the machine learning models by training the model with datasets which are cleaned, processed, transformed, analyzed and best features are selected to fit the model. For different approaches, two classification problems and one regression problem are taken to solve. The Dataset 1 is a classification problem, which has multi class that which player will win the game, either White, Black or the game is Draw. The Dataset 2 is a regression problem, in that price of the used cars will be predicted by using the relevant features of the cars such as manufacturer, color, condition, year of manufacturing etc., The Dataset 3 is also a classification problem, which is about the candidates who applied for Data Scientist position. The prediction will be whether the candidate leaves the job or not, after attending the training in the company.**

Keywords— Chess, price, Data Scientist, Decision tree, Random Forest, training, Winner, Classification

I. INTRODUCTION

In this project, the main objective is to identify best performing model and the steps to proceed that for each dataset. The first dataset is chess game dataset. Data science is all about finding patterns in data, that is why chess has been one of the games that has mostly invested areas of AI in the past. Chess has been tremendously more popular at present and it is getting trend because more people playing at this pandemic time. Winning the match depends on the ratings of the two players, timing and moves taken by both the players. This dataset is the set of more than 20,000 games collected from the users on Lichess.org, which helps to analyze and predict who will be the winner of the game. The second dataset is about used cars. It is the collection of used cars sale in craigslist website within United states. We cannot find whether the car is in good condition and properly working without seeing it in person. The seller can sell the car by higher amount. But using their features of the car, we can predict the exact price for the car. Using this data, the price of the used cars is being predicted. The third dataset is HR Analytics- Job change of Data Scientist. A company wants to hire for the Data Scientist position, who passes some courses will be conducted by the company. Many candidates are signed up for this training program. The company needs to know which candidate need to be stayed after the training program and who wants to move on to new employment after the training conducted by them. This is useful for the company to save the money and consumption of time and also for increasing the standard of training program. This dataset is prepared to predict that who will leave their current job by understanding the factors that affects their decision.

A. DATASET 1: CHESS GAME

This Dataset includes 20,059 rows and 16 columns. The dataset is sourced from Kaggle website. This link to access the dataset is provided below.

https://www.kaggle.com/datasnaek/chess

Research Question:

1. Can we predict the winner of the game? (Classification Problem)

B. DATASET 2: USEDCARS

The Used Cars dataset consists of 44,1397 rows and 25 columns. This dataset is taken from Kaggle website. The link of the dataset to access is provided below.

https://www.kaggle.com/austinreese/craigslist-carstrucks-data

Research Question:

1. Can we forecast the price of the used car? (Regression Problem)

C. DATASET3: HR ANALYTICS-JOB CHANGE OF DATASCIENTISTS

The Job change of Data Scientist dataset contains 19,159 rows and 14 columns. This dataset is used from Kaggle website. The link to access the dataset is provided below.

https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists

Research Question:

1.Can we predict the candidate who change into new employment after the training?

2.What are the features that influence their decision?

II. RELATED WORKS

2.1. DATASET 1: Chess game

This paper [17] explains traditional chess engines extensively investigate progressing chess position possibilities in the board in determining effective moves to play. The author introduced a new technique for playing chess using algorithms in machine learning. The improvement of the evaluation function for endgame positions, the author proposed using an artificial neural network. This outcome is obtained from the three end games. The result shows that the project is better to defeat opponent who gives the good survivability.

All of today's sophisticated two-player programs, according to this analysis [1], have used some variability of both the alpha-beta neural network. The declaration is derived from a pruning method that decreases importantly the game tree expansion and significantly reduces the expanded game tree, allowing for in-depth search. As a result, improving evaluation functions will be useful to create more interesting chess games engine.

The author examined some of the machine learning methods are being used to solve game positions in this research paper [3]. The author had utilized ANN to produce outcome of KING ROOK KING game. The goal of the paper is to forecast the next move of the game.

One of the issues with the current engines' performance is the need to increase the accuracy and complexity of time. The problem is noted, by introducing an online search to offline result estimation process. This idea is changed into action, the offline steps to do is established by some techniques explained in this paper [2].

The chess games are in case, the chess players commonly used particular methods or strategy to win. The final games are instead required many pieces of coins, which is not that easy to solve. The pieces in the games involves many numbers of possible move as per configuration is mentioned in this paper [16].

2.2 DATASET 2-Used Cars

For this study, the author performed a study on the outputs of linear-regression depends on algorithms in machine learning[15]. Every model from German e- commerce website is data of used cars. The results are Mean absolute error is 0.28, gradient boosting regression trees are well performed. By following that, 0.35 MSE for Random forest and 0.65 for linear regression

It is explained in this paper[18] about developing a method for predicting used car in linear regression, Xgb boost and random forest. Every method depends on data obtained from an e-commerce site. The main thing of this paper is to find the good fit model among all other methods to detect the price of used cars.

The author in this study [9] conducted a performance study of regression models. The data in this paper is extracted from e-commerce website and then changed according to the project. The result is contained with 304,133 rows and 11 columns. The researchers used the data to check specific Dataset using algorithms. Therefore, every model is calculated using same test data. The mean absolute error is a criterion to check and compare the results. The best model is achieved with gradient boosted regression trees, of MAE Value if 0.28. Random forest regression is of 0.35 MAE value. The result concludes the usage gradient boosted regression tress to find the model development to predict the price in this paper [9].

The evaluation and results are determined in this paper [10], using test data as input into the multiple linear regression model, gradient boosted and random forest regression. The mean absolute error is a criterion to check and compare the models. The MSE of 0.28, is the best fit gradient boosted regression model. Random forest regression gets 0.35 of MSE. By comparing these two, MSE is higher in linear regression of 0.5.

2.3 DATASET 3- A Job Change of Data Scientists

According to the findings [12], that is not necessary for management to agree on a specific resume even before considering position. By staying same position, some higher professional skilled person, while others prioritize working experience and industry knowledge.

In today's job market, comprehending these recruiting patterns is just becoming increasingly important. Relevant job search engines return resumes that are relevant to the keywords entered. The difficulty in selecting the best profile in the paper increases with an increase of search results from these search results increases[5].

According to the paper [14], the concept of Human Resource (HR) personnel in hiring and looking profiles is highly important. The profiles will be created by HR and manually arranged by ranking. The scheme detects the intelligence in the hidden patterns. In addition to the traditional search technique, highly trained models predict the ranking of highest accuracy.

The cosine distance and input questions are estimated using R package. The model is calculated using venn diagram and its results [7].

III. METHODOLOGY

The process of finding helpful information from a data collection is called as Knowledge Discovery in Databases (KDD). This strategy utilizes data preprocessing and selection, data cleansing, integrating previous understanding of data sets, and evaluating precise prediction from measured data. The steps associated with the process of KDD are,

- The data is extracted from kaggle website and dependent variable is selected.
- Data Cleaning is cleaning the null values either by removing it or using imputation to fill it. The outliers are checked for numerical variables and it is handled by removing it.
- The data is studied and understood the type of data and the statistical information about the data to perform good analysis to evaluate the result.

- Data transformation is being performed and Variable selection is done for input the good variables into the model for using algorithm and to interpret the results.
- The findings were captured and reported.

Knowledge Discovery in Databases (KDD) Methodology is used in this project for Data Mining. The steps have performed to analyze the data are performed below.

1.DATASELECTION

The Three Dataset that we selected are from three different fields such as

1.Chess game

2.Used Cars

3.A Job change of Data Scientist

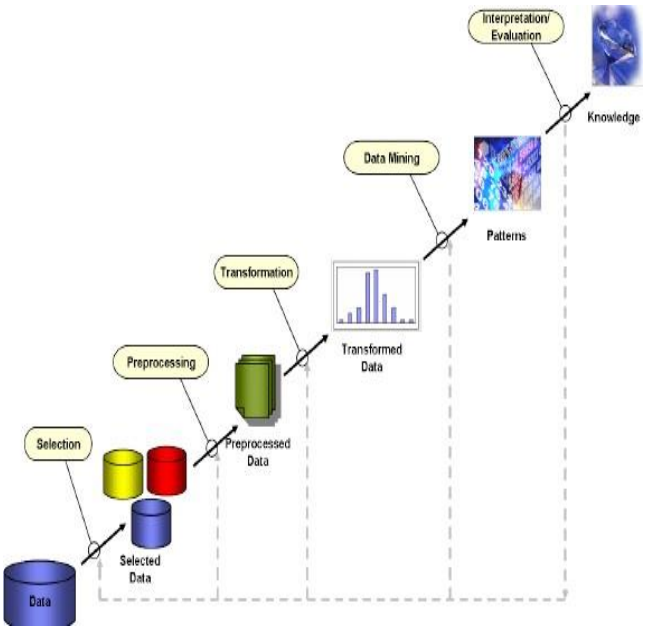

Fig-1 Process of KDD (Knowledge Discovery in Databases)

2.DATA PRE-PROCESSING*:*

2.1. Dataset-1: Chess game

A. Data Cleaning

This Dataset contains feature variables such as id,created at, rated, turns, last move at, victory status, white id, increment code, white rating, black id, opening eco, black rating, moves, opening name, opening ply and target is winner. This is shown in the figure-1.The target variable is polytomous, which has three classes such as white, black and draw.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20058 entries, 0 to 20057
Data columns (total 16 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   id              20058 non-null  object
 1   rated           20058 non-null  bool
 2   created_at      20058 non-null  float64
 3   last_move_at    20058 non-null  float64
 4   turns           20058 non-null  int64
 5   victory_status  20058 non-null  object
 6   winner          20058 non-null  object
 7   increment_code  20058 non-null  object
 8   white_id        20058 non-null  object
 9   white_rating    20058 non-null  int64
 10  black_id        20058 non-null  object
 11  black_rating    20058 non-null  int64
 12  moves           20058 non-null  object
 13  opening_eco     20058 non-null  object
 14  opening_name    20058 non-null  object
 15  opening_ply     20058 non-null  int64
dtypes: bool(1), float64(2), int64(4), object(9)
memory usage: 2.3+ MB
```

Fig-2 Structure of Dataset 1

```
how many nulls are there

id                0
rated             0
created_at        0
last_move_at      0
turns             0
victory_status    0
winner            0
increment_code    0
white_id          0
white_rating      0
black_id          0
black_rating      0
moves             0
opening_eco       0
opening_name      0
opening_ply       0
dtype: int64
```

Fig-3 Null Values of Dataset-1

In the Figure-3, the null values of dataset 1 can be seen. It shows that there is no null values in this dataset.

Feature Engineering:

The unnecessary and unrelated columns in this dataset have removed by using drop function. The Feature Engineering is encoding with 0s and 1s to the categorical variables. The variables such as rated, victory status and winner are encoded with 0s and 1s by replacing the categories.

Correlation:

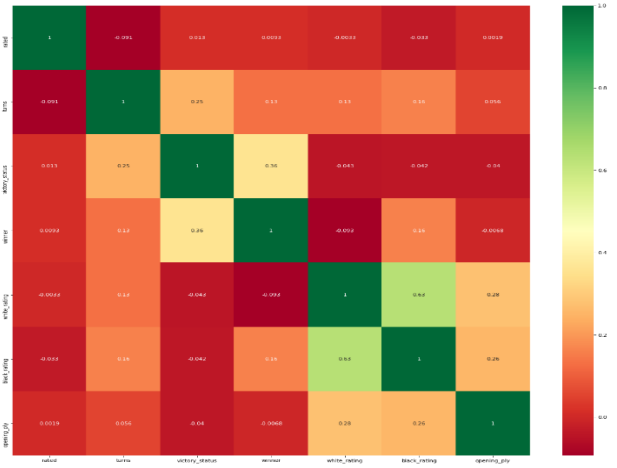

Fig-4 Correlation for dataset 1

In the Figure-4, heatmap that victory status and turns has strong correlation with winner. The correlation of victory status and turns are 0.36 and 0.16. So, it will be dropped the features apart from this. Multicollineraity means that the independent variables need to be no relation with each other. It is checked between each variable. The variable turns and victory status has high positive correlation of 1 between them. This problem is called Multicollinearity problem. Either one variable should be removed before applying it into the model. Turns- variable is removed from the dataframe before applying the algorithm. The unique values of categorical variables are described using value_counts function.

B. Variable Selection:

The selected feature variables from dataset are rated, victory status, white rating, black rating, and opening ply, these are used to predict the target variable-winner.

2.2. Dataset-2: Used Cars

A. Data Cleaning

This Dataset consists of 17 features, in which price is the target variable and the 16 features.The structure of data is seen in the Figure 5.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 441396 entries, 0 to 441395
Data columns (total 17 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   region        441396 non-null  object
 1   price         441396 non-null  int64
 2   year          440359 non-null  float64
 3   manufacturer  423019 non-null  object
 4   model         436057 non-null  object
 5   condition     257554 non-null  object
 6   cylinders     253231 non-null  object
 7   fuel          438515 non-null  object
 8   odometer      437018 non-null  float64
 9   title_status  432451 non-null  object
 10  transmission  438769 non-null  object
 11  drive         307747 non-null  object
 12  size          125812 non-null  object
 13  type          346047 non-null  object
 14  paint_color   308374 non-null  object
 15  county        0 non-null       float64
 16  state         441396 non-null  object
dtypes: float64(3), int64(1), object(13)
memory usage: 57.2+ MB
```

Fig-5 Structure of Dataset 2

```
region              0
price               0
year             1037
manufacturer    18377
model            5339
condition      183842
cylinders      188165
fuel             2881
odometer         4378
title_status     8945
transmission     2627
drive          133649
size           315584
type            95349
paint_color    133022
county         441396
state               0
dtype: int64
```

Fig-6 Missing Values of Dataset 2

HANDLING MISSING VALUES:

The missing values in Dataset 2 are larger in numbers. So, the columns with more than 55% of missing values are to be removed. The categorical variables are handled by filling the frequent values in the column. The unique values are viewed by using value count function. The rest of the numeric variable columns are dropped from the data-frame. The Figure-7 represents that the variables after removing all the missing values.

```
region          0
price           0
year            0
manufacturer    0
model           0
condition       0
cylinders       0
fuel            0
odometer        0
title_status    0
transmission    0
drive           0
type            0
paint_color     0
state           0
dtype: int64
```

Fig-7 No null values

OUTLIERS:

The Outliers are the extreme values such as minimum extreme and maximum extreme values. In this dataset, the price variable's outliers are checked. The minimum extreme value is 0 and maximum extreme is 3,736,928,711. The price variable can never be 0. So, the values below and equal to 0 are to be removed. The Figure-8 shows the outliers of the price variable.

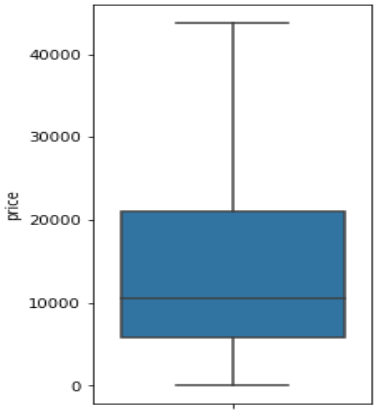

Fig-8 Outliers of price

HANDLING THE OUTLIERS:

The Outliers of price are removed with IQR (Interquartile Range). The interquartile range of Q1 and Q3 are taken between 0.05% and 0.95%, in which below 0.05% and above 0.95% detected. The detected values are identified as False and other values are identified as True as well. Now, the False values are removed from the original dataframe. This can be seen on the figure 9.

```
0          True
1          True
5          True
11         True
12         True
          ...
441349     True
441367    False
441374    False
441383    False
441384     True
Name: price, Length: 111624, dtype: bool
```

Fig-9 Identification of Outliers

The Outliers are also checked in the other numeric variable- odometer. It is little in amount, so that will not affect the model.

Label Encoding:

The categorical variables have many categories. Label encoding is encoding the categories with numerical values.

MinMax Scaler:

The odometer is a variable, which has large magnitude, reducing the measure of the variable to protect the dominance of prediction model. In order to, maintain all the variables in the same level, MinMaxScaler is applied to predict the model perform better.

Variable Selection:

The three methods are for selecting variables in Multiple Linear Regression and the one with higher accuracy is selected.

1.Filter Method:

The filter method is filtering and taking subset of the related variables. The filtering method is a method by using correlation matrix with Pearson correlation in the figure 10.

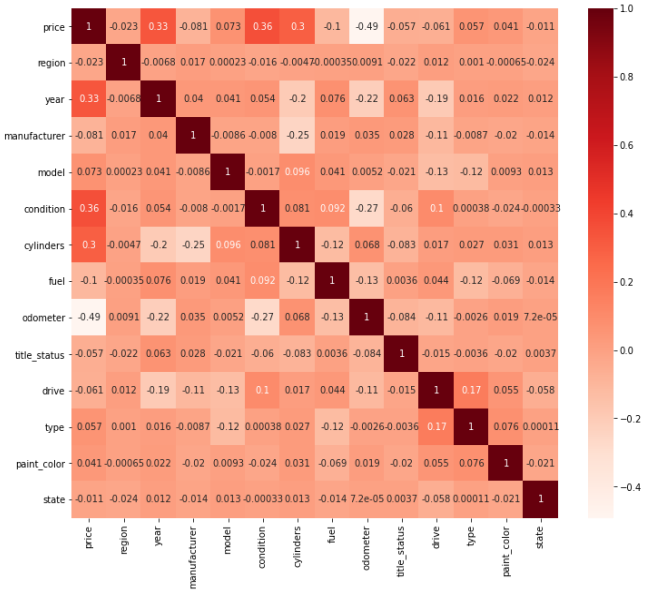

Figure-10 The Correlation of dataset 2

The heatmap that odometer, fuel,transmission, cylinders and year have strong positive and negative correlation with price. The correlation of odometer, cylinders, transmission and year are -0.1, -0.49, 0.3, 0.36, and 0.34. It will be dropped the features apart from this. Multi-Collineraity means that the independent variables need to be uncorrelated with each other. It is checked between each variable and there is no correlation between independent variables. So, there is no multi-collinearity problem.

The data is splitted randomly with test data 25%. The training prediction variables are 75202 rows, the training independent variables are 75202 rows, the testing prediction variables are 25068 rows, and the testing dependent variables are 25068 rows are obtained. The regression model runs with Pearson correlation method. The training data is fitted into model and the coeffiicients are obtained for fuel, odometer, cylinders, transmission and year are -3207.24204,-115508.113976, 2790.046702, 3958.984368, 346.101467 respectively, which can be seen in the figure-10.

```
Intercept : -676929.393312779

        features    coeficients
0           fuel   -3207.242040
1       odometer -115508.113976
2      cylinders    2790.046702
3   transmission    3958.984368
4           year     346.101467
```

Fig-11 The intercept and coefficients for filtering technique

The R square score is **0.49754096786136004**.

2.BACKWARD ELIMINATION (Wrapper Method):

This is a process, which produces good results. All the features are input into model. The evaluation of model is to be performed and worst variables are removed., until good model comes. The performance metric used to evaluate feature performance in this case is p value. If the p value is greater in value of 0.05, the feature is removed; otherwise, it is retained. The final dataset after being p-value is greater than 0.05 has removed are applied into the model. In Figure-12, the p-values of all the variables are shown.

```
const           0.000000e+00
region          6.669065e-09
year            0.000000e+00
manufacturer    2.791918e-03
model           1.070952e-40
condition       2.212720e-13
cylinders       0.000000e+00
fuel            0.000000e+00
odometer        0.000000e+00
title_status   1.354399e-221
transmission    0.000000e+00
drive          4.891557e-239
type            6.713421e-61
paint_color     1.544682e-31
state           4.931075e-34
dtype: float64
```

Fig-12 P-values

The data is splitted randomly with test data 25% by selected variables. The training prediction variables are 75202 rows, the training independent variables are 75202 rows, the t

esting prediction variables are 25068 rows, and the testing dep endent variables are 25068 rows are obtained. The regression model runs with Pearson correlation method. The training data is being fitted into model and the coefficients are obtained for all the variables are shown in the figure 13.

```
Intercept : -634150.1030364304

        features    coeficients
0         region      -1.208449
1           year     325.116516
2   manufacturer       7.149954
3          model       0.101814
4      condition     143.786637
5      cylinders    2710.873980
6           fuel   -3094.885780
7       odometer -120243.790049
8   title_status    -778.827707
9   transmission    3914.726621
10         drive   -1007.531793
11          type      87.362617
12   paint_color      59.858262
13         state     -18.102817
```

Fig-13 output coefficient of regression for backward coefficients

The R square score is **0.5103226460612391**.

1. Recursive Variable Elimination:

This method removes variables and builds a model depends on those that remain. It gives accuracy to rank the features in order of importance. The score of 14 variables obtained as 0.510323. The 14 features are fitted into the model. The data is splitted randomly with test data 25%. The training prediction variables are 75202 rows, the training independent variables are 75202 rows, the testing prediction variables are 25068 rows, and the testing dependent variables are 25068 rows are obtained. The regression model runs with Pearson correlation method. The training data is being fitted into model and the coefficients are obtained for all the variables are shown in the figure-12.

```
Intercept : -636888.001334139

        features    coeficients
0         region      -1.202529
1           year     326.759272
2   manufacturer       7.177594
3      cylinders    2744.438448
4           fuel   -3061.689327
5       odometer -120492.440518
6   title_status    -780.234618
7   transmission    3956.056734
8          drive   -1041.265072
9           type      81.714451
10   paint_color      61.512469
11         state     -18.005108
```

Fig-14 Intercept and Coefficients

The R square score is **0.5092248866079689.**

2.3. Dataset 3: A Job change of Data Scientist

A. Data cleaning

This Dataset contains feature variables and target variable, this is seen in the figure-14. The target variable is polytomous, which has three classes such as white, black and draw.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19158 entries, 0 to 19157
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   enrollee_id             19158 non-null  int64
 1   city                    19158 non-null  object
 2   city_development_index  19158 non-null  float64
 3   gender                  14650 non-null  object
 4   relevent_experience     19158 non-null  object
 5   enrolled_university     18772 non-null  object
 6   education_level         18698 non-null  object
 7   major_discipline        16345 non-null  object
 8   experience              19093 non-null  object
 9   company_size            13220 non-null  object
 10  company_type            13018 non-null  object
 11  last_new_job            18735 non-null  object
 12  training_hours          19158 non-null  int64
 13  target                  19158 non-null  float64
dtypes: float64(2), int64(2), object(10)
memory usage: 2.0+ MB
```

Figure-15 The Structure of Dataset 3

The nulls values of the dataset 3 are shown in the figure 15.

```
how many nulls are there

city                      0
city_development_index    0
gender                 5016
relevent_experience       0
enrolled_university     417
education_level         512
major_discipline       3125
experience               70
company_size           6560
company_type           6774
last_new_job            463
training_hours            0
target                 2129
enrollee_id           19158
dtype: int64
```

Figure-16 The Null Values

The missing values of the variables in percentage represents in the figure 16. These null values are dropped from the data. Because it does not affect the target variable.

	Missing Values	% of Total Values
enrollee_id	18014	89.9
company_type	6055	30.2
company_size	5887	29.4
gender	4306	21.5
major_discipline	2471	12.3
target	2018	10.1

Fig-17 Null values percentage

All other variables such as gender, education level, enrolled university, experience, last new job, target are to be filled with frequent values in the categories.

Label Encoding:

The categorical variables are encoded from categories into numeric. This is called Label encoding.

Correlation:

The heatmap that city_development_index, education_level and experience have high positive and negative correlation with target. So, features aredrpped apart from this.

Multicollineraity means that independent variables are not relation to each other. city_development_index and experience has strong correlation between each other. multicollinearity is occurred.So remove the two variables before building the model. This can be seen in the correlation figure-17.

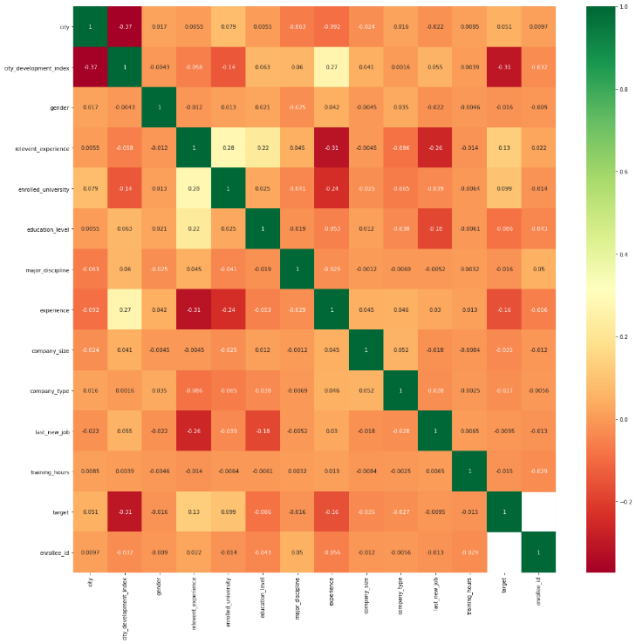

Figure-18 Correlation of Dataset 3

IV. EVALUATION AND RESULTS

4.1 DATASET 1: CHESS GAME:

Model 1:

Random Forest:

The Random Forest algorithm is applied in this model.

Prediction and Analysis:

The Random Forest Classifier model has the accuracy of **0.6612662013958126**.

```
RESULT
Random Forest Model Acc :  0.6612662013958126
```

Figure-18 Accuracy of Random Forest Classifier

The confusion matrix:

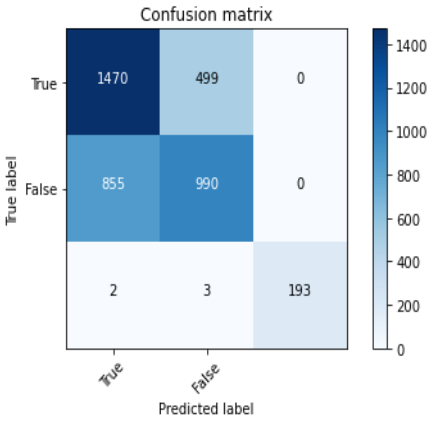

Figure-19 confusion matrix

The output results are shown in the figure 20. The F1 score is **0.75**, which is a good score that it means the model is fitted well.

	precision	recall	f1-score	support
0	0.63	0.75	0.68	1969
1	0.66	0.54	0.59	1845
2	1.00	0.97	0.99	198
accuracy			0.66	4012
macro avg	0.77	0.75	0.75	4012
weighted avg	0.66	0.66	0.66	4012

Fig-20 RESULTS

Model 2:

XGB BOOST CLASSIFIER:

The XGB Boost algorithm is applied in this model.

Predictive and Analysis:

The XGB Boost Classifier model has the accuracy of **0.665**.

```
RESULT
XGBoost Model Acc :  0.6650049850448654
```

Figure-21 Accuracy of XGB Boost Classifier

The confusion matrix:

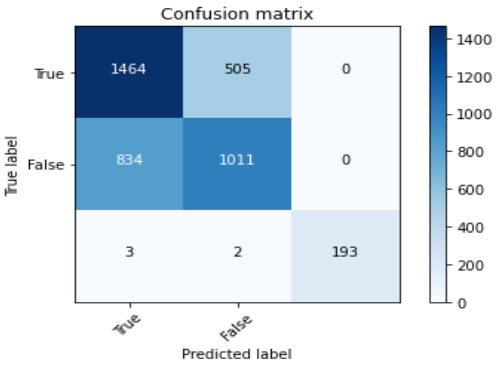

Figure-21 confusion matrix

The output results are shown in the figure 20.
The F1 score is **0.76**, which is a good score that it means the model is fitted well.

	precision	recall	f1-score	support
0	0.64	0.74	0.69	1969
1	0.67	0.55	0.60	1845
2	1.00	0.97	0.99	198
accuracy			0.67	4012
macro avg	0.77	0.76	0.76	4012
weighted avg	0.67	0.67	0.66	4012

Fig-22 RESULTS

4.2 DATASET 2: Used Cars

Model 1:

Decision Tree:

Decision Tree is applied in this model.
Prediction and analysis:
The R square value is **0.49754096786136004.**

```
score = r2_score(y_test, lr_pred_p)
score
```

```
0.49754096786136004
```

Fig-23 Output of Decision Tree

Model 2:

Decision Tree Regressor:

Decision Tree Regressor is applied in this model.
Prediction and analysis:
The Rsquare value is **0.5092248866079689.**

```
score = r2_score(y_test, lr_pred_rfe)
score
```

```
0.5092248866079689
```

Fig-24 Output of Decision Tree Regressor

4.3. DATASET 3: A Job change of Data Scientist.

Model 1:

XGB Boost:

XGB Boost is applied in this model.

Prediction and analysis:

The Rsquare value is **0.7785507246376812.**

```
RESULT
XGBoost Model Acc :  0.7785507246376812
```

Fig-25 Output of XGB Boost

Confusion Matrix:

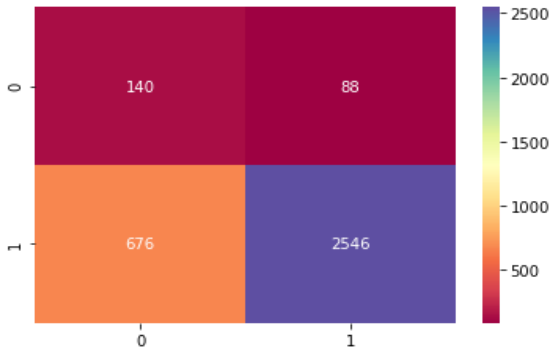

Fig-26 Xgb Boost

ROC CURVE:

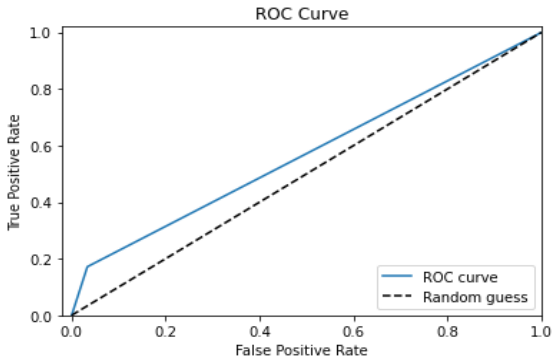

Fig-27 ROC Curve of XGB Boost

The F1 Score is **0.73,** which is good score for the model.

	precision	recall	f1-score	support
0.0	0.79	0.97	0.87	2634
1.0	0.61	0.17	0.27	816
accuracy			0.78	3450
macro avg	0.70	0.57	0.57	3450
weighted avg	0.75	0.78	0.73	3450

Fig-28 Output

Model 2:

Random Forest:

Random Forest is applied in this model.
Prediction and analysis:
The R square value is **0.7794202898550725**.

```
RESULT
Random Forest Model Acc :  0.7794202898550725
```

Fig-9 Output of Random Forest

Confusion Matrix:

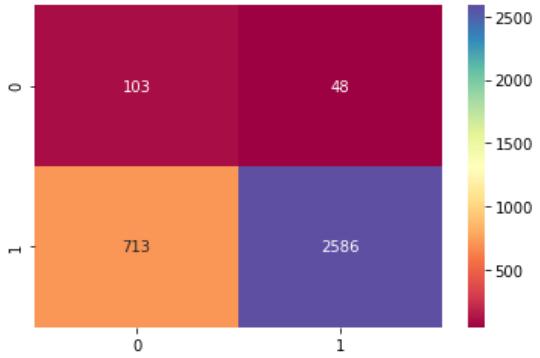

Fig-29 Random Forest

ROC CURVE:

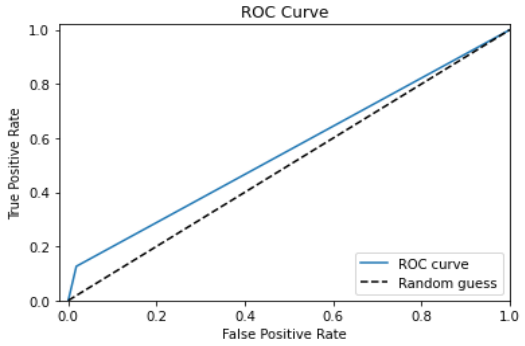

Fig-30 Random Forest

The F1 Score is **0.72,** which is good score for the model.

	precision	recall	f1-score	support
0.0	0.78	0.98	0.87	2634
1.0	0.68	0.13	0.21	816
accuracy			0.78	3450
macro avg	0.73	0.55	0.54	3450
weighted avg	0.76	0.78	0.72	3450

Fig-31 Output of Random Forest

Model 3:

KNN MODEL:

KNN algorithm is applied in this model.

```
RESULT
KNN Model Acc :  0.7892753623188405
```

Fig-32 Output of KNN

Confusion Matrix:

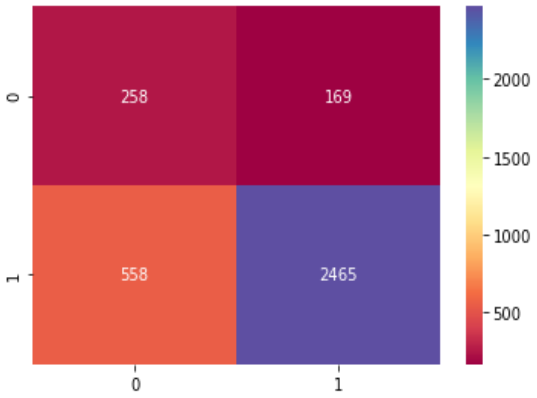

Fig-34 K- nearest neighbour

ROC Curve:

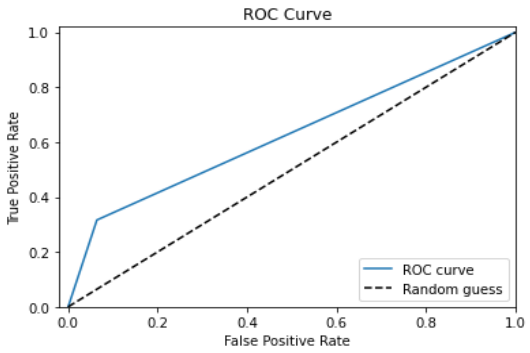

Fig-35 KNN

The F1 score is **0.76**, which is a better score compared to the two models.

	precision	recall	f1-score	support
0.0	0.82	0.94	0.87	2634
1.0	0.60	0.32	0.42	816
accuracy			0.79	3450
macro avg	0.71	0.63	0.64	3450
weighted avg	0.77	0.79	0.76	3450

Fig-36 Output of KNN

V. Conclusion

The Data are extracted from Kaggle, Pre-processed and transforming the data to interpret data mining and evaluate the best results from fitting various models in machine learning techniques. In Dataset1, Both Random Forest, XGB Boost Classifier proved to be the best model with accuracy of 66% and 66% respectively and F1 score are 0.75 and 0.76 respectively, for predicting which player would win. In Dataset 2, comparing to Random Forest, Random Forest regressor proved to be the best model with R square 0.50, for predicting the price of used cars. In Dataset 3, Compared to XGB Boost and Random Forest, K-Nearest Neighbour

algorithm has proved to be the best fitted model with accuracy of 78% and F1 score 0.76.

For Future works, the advanced techniques will be used to proceed with best and efficient evaluations and results.

REFERENCES

[1] A. Hauptman and M. Sipper. Using genetic programming to evolve chess endgame players. In Proceedings of the 8th European Conference on Genetic Programming, pages 120– 131. Springer, 2005.

[2] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, pp. 1–16, jan 2007

[3] D. DeCoste. The signifcance of kasparov vs deep blue and the future of computer chess. ICCA Journal, (21):33–43, 1998.

[4] G. Haworth and M. Velliste. Chess endgames and neural networks. ICCA journal, 21(4):211–227, December 1998.

[5] G.Rossum, "Python Reference Manual," Amsterdam, The Netherlands, The Netherlands, Tech. Rep., 1995

[6] H. Berliner. Construction of evaluation function for large domains. Artificial Intelligence, 99:205–220, 1992.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vander-(5)plas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duches-nay, "Scikit-learn: Machine Learning in fPgython," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011

[8] J. Si and R. Tang. Trained neural network play chess endgames. IJCNN, 6:3730, 1999.

[9] J. Morgan, "Classification and Regression Tree Analy-sis," Bu.Edu, no. 1,p.16,2014.[Online].Available:http://www.bu.edu/sph/files/2014/05/MorganCART.pdf

[10] Junjie Wu, Advances in K-means Clustering, Springer-Verlag Berlin Heidelberg, 2012.

[11] Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, Mining of Massive Datasets, Stanford Infolab, 2014.

[12] Kumar, A., Pandey, A., & Kaushik, S. (2017). *Machine Learning Methods for Solving Complex Ranking and Sorting Issues in Human Resourcing. 2017 IEEE 7th International Advance Computing Conference (IACC).* doi:10.1109/iacc.2017.0024

[13] M. Auton'es, A. Beck, P. Camacho, N. Lassabe, H. Luga, and F. Scharffe. Evaluation of chess position by modular neural network generated by genetic algorithm. EuroGP, pages 1–10, 2004.

[14] Michael Steinbach, Vipin Kumar, Pang-Ning Tan, Introduction to Data Mining, Pearson Publications, 2006.

[15] Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., & Boonpou, P. (2018). Prediction of prices for used car by using regression models. 2018 5th International Conference on Business and Industrial Research (ICBIR). doi:10.1109/icbir.2018.8391177.

[16] N. Lassabe, S. Sanchez, H. Luga, and Y. Duthen. Genetically programmed strategies for chess endgame. In GECCO '06: Proceedings of the 8th annual conference on Genetic and evolutionary computation, pages 831–838, New York, NY, USA, 2006. ACM Press.

[17] Samadi, M., Azimifar, Z., & Jahromi, M. Z. (2007). Learning: An Effective Approach in Endgame Chess Board Evaluation. Sixth International Conference on Machine Learning and Applications (ICMLA 2007). doi:10.1109/icmla.2007.48

[18] S. Peerun, N. H. Chummun, and S. Pudaruth, "Predicting the Price of Second-hand Cars using Artificial Neural Networks," The Second International Conference on Data Mining, Internet Computing, and Big Data, no. August, pp. 17–21, 2015

[19] S. Pudaruth, "Predicting the Price of Used Cars using Machine Learning Techniques," International Journal of Information & Computation Technology, vol. 4, no. 7, pp. 753–764, 2014

[20] Yanchang Zhao, R and Data Mining: Examples and Case Studies, 2013.

.