Potential Performance Predictor - P^3 - Group 19

Subodh Gujar
sgujar@ncsu.edu

Deep Mehta
dmmehta2@ncsu.edu

Kartiki Bhandakkar
kbhanda3@ncsu.edu

1 BACKGROUND

Problem Statement

Hiring an employee who would be a right fit for the company has always been a challenge for the human resources department at almost every company on this planet. Traditional approaches included manually evaluating an employee after many rounds of interviews and based on feedback from interviewers, a decision was made as to whether a particular employee would be an ideal fit within the company or not. An ideal fit is defined as someone who understands the company's vision, thrives in the work environment, and is a long-term employee. Attrition has been one major concern for these companies recently, as many employees nowadays tend to switch companies often which in turn leads to a loss of investments the company made in employee skill development.

This project aims to build a machine-learning model/s that can predict employee performance during the hiring process. The model will consider existing employees in the company and utilizing their performance over the last years, evaluate how a new employee with similar attributes will fit into the work environment at that company. This approach is unique in the sense that each person responds differently to various aspects of a work environment, and if we utilize the fact that an existing employee in that particular role has been performing well in that work environment, then it would be more accurate rather than just comparing skills or education. A major ethical issue that we will be addressing as part of this project is to look into attributes that may lead to bias and make decisions to remove any sort of bias during the hiring process.

Related Work

Several studies in the past have used data mining for extracting rules and predicting certain behaviors for the employee performance prediction use case. Some of these papers are referenced to understand what worked well in the past and the major concerns.

Researchers like Chein and Chen (2006) [1] has worked on the improvement of employee selection, by building a model, using data mining techniques, to predict the performance of new applicants. One interesting point here was they excluded age, gender, and marital status to get rid of any bias in the employee hiring process. The following observations were made: Degree, School Tire and Job Experience highly affected the job performance of employees.

Kahya [2] made an interesting discovery while researching the effects of working conditions and environment on job performance. The research showcased that working conditions had both positive and negative impacts on performance. While, highly educated and qualified employees demonstrate low performance in bad environments, interestingly, employees with low qualifications demonstrate high performance despite bad environments.

Qasem and Eman [5] utilized the CRISP-DM methodology (Cross Industry Standard Process for Data Mining) (CRISP-DM, 2007) to make performance predictions of employees. They utilized two techniques, Decision Trees (ID3 and C4.5) and Naive Bayes, with 10-fold CV and hold out. Their work concluded that job title had the most impact on employee performance, but job satisfaction and work environment had a slight effect on the employee's performance.

The authors Mosavi, A., Sajedi Hosseini, F., Choubin [4] investigated the use of ensemble learning techniques for predicting groundwater potential and compares the performance of three ensemble methods with four individual machine learning models. The study finds that ensemble methods, especially AdaBoost, are effective in predicting groundwater potential and can be useful for decision-making.

The authors Kotsiantis, Sotiris & Pintelas, P. [3] propose a hybrid ensemble method that combines bagging and boosting techniques for classification tasks. The method creates multiple subsets of the training data using bagging and trains a weak classifier for each subset, which is then combined using boosting to form a strong classifier. The proposed method outperforms both bagging and boosting and is competitive with other state-of-the-art ensemble methods, according to the experiments conducted by the authors.

2 METHODS

Novel Aspects

The novel aspects of this project are :

(1) Performing exploratory data analysis of dataset to understand feature co-relation and experiment with different features to identify bias in the model and remove it.
(2) Combining different machine learning models, such as through ensemble methods, to increase accuracy and improve the performance of the overall model.
(3) Considering the use case, this dataset will also be used to draw a meaningful hypothesis, that goes beyond simple 'yes' and 'no'.

Approach

(1) Firstly, we analyze the dataset manually and utilize domain knowledge to make assumptions as to what candidates are suitable for the target column for supervised learning.
(2) Next, we would get rid of unnecessary features that would have no impact on the predictions to be made using a correlation matrix.
(3) To overcome the imbalance of different classes represented by data for target columns, we will apply Smote technique.
(4) Various feature selection techniques will be applied, and the one which gives the best accuracy will be selected. Currently, we are planning on checking PCA, Lasso Regression, and Select K-Best to reduce features and thus simplify the models.
(5) Predicting job satisfaction and environment satisfaction and use these predictions as features for predicting attrition using various ML techniques including Random forest, Logistic regression, Adaboost, KNN, Naive Bayes, SVM, and MLP Neural Network.
(6) Create an Ensemble of seven ML techniques, Random forest, Logistic regression, Adaboost, KNN, Naive Bayes, SVM, MLP Neural Network, and Decision Tree and write custom logic to implement boosting and train the model by running it for N iterations.
(7) Apply K-Means clustering to understand how data is perceived and divided among clusters and see if any particular feature may be creating bias or not.
(8) Identify bias and draw meaningful hypotheses by data analysis and interpretation of graphs.
(9) Remove biased features from the Ensemble models and retrain the model.

Rationale

As an initial implementation, we decided to go with Smote technique to overcome the class imbalance in the dataset, because this would help in reducing bias towards a particular class of target variables.

There are over 20+ features in the dataset, and utilizing all of these would result in over-fitting of the model to training data. Thus, feature selection would help in this case because we are interested in identifying the features which are most relevant for predicting our outcome variable. We will be comparing various feature selection techniques on this particular dataset and identify the one which would give the best accuracy.

Job satisfaction and Environment satisfaction didn't have a promising accuracy, so predicting those independently for a future employee might not be the best choice. But interestingly, the accuracy of attrition improves if job satisfaction and environment satisfaction are utilized as features, so we decided to first predict those values and then utilize those as features to predict attrition. The reason why these need to be predicted is that for a new employee, we won't have any data related to job satisfaction and environment satisfaction, so we assume that if these can be predicted and then, in turn, be used as features for predicting attrition, that would help in increasing accuracy.

The purpose of evaluating multiple machine learning techniques is to come up with a set of preferable techniques that will be utilized while using the Ensemble method to make predictions.

Utilizing multiple ML techniques and writing a custom logic for boosting will be beneficial to get a better accuracy overall and also avoid over-fitting to a particular dataset for a particular model. This would run N iterations across various models and based on the output, get the majority of all the models to calculate the error, alpha, and weights for the next iteration.

Clustering might be beneficial in this case to see how data is split into various clusters and analyze it to see if any potential biased features exist on which clusters might be created.

Analyzing graphs of potentially biased features against attrition in various circumstances and identifying if any particular feature causes bias against a particular section of society and if it does such features must be avoided to ensure an unbiased and fair model is created.

Finally, removing the biased features, and retrain the model and that would be the final unbiased and accurate model to be used to evaluate the potential employees of a company.

3 PLAN & EXPERIMENTS

Datasets

IBM HR Analytics Employee Attrition & Performance (Link).

This is a fictional data set created by IBM data scientists. This dataset describes 35 attributes related to 1500 employees across the company.

This dataset provides relatable attributes like Department, Education, Environment Satisfaction, and many more, that allowed us to make appropriate predictions on the performance of the employee as described in the project idea above.

Hypotheses

The following hypotheses are made before running the experiments:

(1) If supervised learning is used, which of the following three columns or their combinations provide the most accurate results: Attrition, Environment Satisfaction, or Job Satisfaction?
(2) Which features would be most effective in predicting performance? And which features would be the least effective and can get rid of?
(3) Does the dataset represent all classes equally? If not, how to overcome this hurdle? And will this create a bias if not resolved?
(4) Which combination of supervised techniques should be implemented in Ensemble such that it yields high accuracy?
(5) Do the attributes Gender, Marital Status, Age, and Distance from home, create a form of bias, and should these be avoided?
(6) What meaningful hypothesis can be derived from this data?

Experimental Design & Description

(1) The dataset is split into test and training data using stratified random sampling with a division of 20% test and 80% training. The test data is utilized at a later stage to predict the accuracy of trained models.
(2) Data normalization has been performed using the Standard-Scalar function to transform features to be on a similar scale.
(3) Made use of confusion matrix to calculate precision, recall, and F1-score using Random forest, Logistic regression, Adaboost, KNN, Naive Bayes, SVM, MLP Neural Network for all three target columns, that are Attrition, Environment Satisfaction, and Job Satisfaction.
(4) Compared various feature selection techniques including Select K-Best, PCA, and Lasso Regression, and selected the best technique which yields better accuracy for this particular dataset, i.e. Select K-Best.
(5) The dataset is analyzed for the division of classes for target columns, and as the classes are unevenly distributed, Smote technique is being utilized to overcome this.
(6) Finally, Smote, Select K-Best, and various machine learning techniques are utilized and results were analyzed to select the models to be used in Ensemble.
(7) An ensemble of seven ML techniques, Random forest, Logistic regression, Adaboost, KNN, Naive Bayes, SVM, MLP Neural Network and Decision Tree are developed based on the results observed in the previous experiments. The prediction is that this would provide higher accuracy and will better fit overall data and not just the current training dataset.

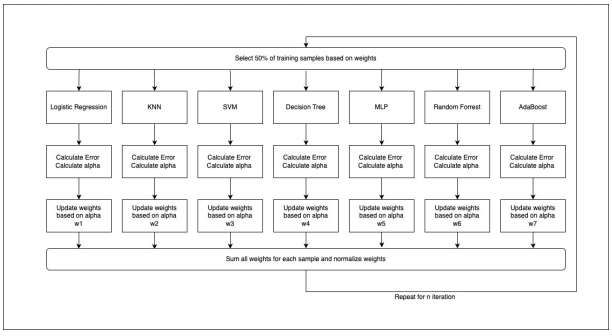

Figure 1: Ensemble architecture

(8) The Ensemble implements logic similar to boosting, where in each iteration the models predict the target variable, and then the majority vote is taken, alpha and error are calculated, and the weights are reassigned for the next iteration. N such iterations are run and finally, the model is trained.

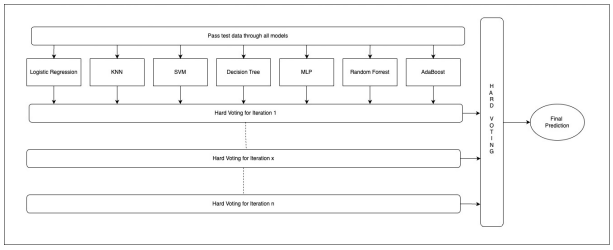

Figure 2: Ensemble iterations

(9) Clustering is applied and experiments are run to analyze if unsupervised learning benefits such a dataset or not. We consider that this will help group people with similar attributes and predict if the new employee will fit in or not. This will also help in identifying if any particular feature is acting as a bias for creating the clusters or not.
(10) Both K-Means and Hierarchical clustering is implemented and results are analyzed and compared.
(11) Generated graphs to compare various features and analyze if a particular feature is creating a bias against a particular section of society or not.
(12) Running experiments with and without the bias columns that are Gender, Marital Status, Age, and Distance from home, and evaluating if these features generate biased outcomes or not.
(13) Finally, biased features are removed and the unbiased Ensemble of models is retrained to predict if a new employee will fit in the company or not.

4 RESULTS

Result

The following graph depicts the distribution of classes in target columns in the given dataset:

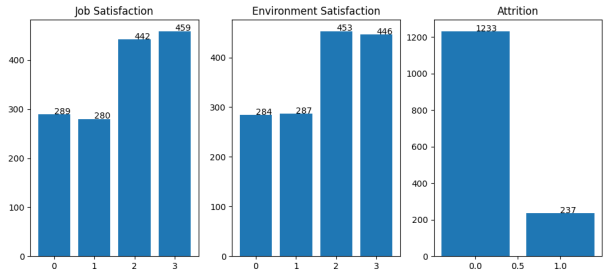

Figure 3: Distribution of classes represented by target variables in dataset

Feature Selection

The below table compares the accuracy obtained after applying PCA, Lasso Regression and Select K-Best and using those filters to train Random Forest classifier.

Feature Selection	Accuracy	F1-Score
PCA	0.82	0.82
Lasso Regression	0.89	0.90
Select K-Best	0.90	0.90

While Lasso and K-Best showed similar results, we decided to go ahead with Select K-Best. Below figure is the confusion matrix of Random Forest model after applying feature selection using Select K-Best technique.

```
Accuracy Score:  0.8978930307941653
Confusion Matrix:
 [[280  25]
 [ 38 274]]
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.92      0.90       305
           1       0.92      0.88      0.90       312

    accuracy                           0.90       617
   macro avg       0.90      0.90      0.90       617
weighted avg       0.90      0.90      0.90       617
```

Figure 4: Select K-Best Accuracy - Random Forest

Below are the features reduced after applying Select K-Best with Attrition as target column:

```
Selected Features Index(['Age', 'BusinessTravel', 'DailyRate', 'Department', 'DistanceFromHome',
       'Education', 'EducationField', 'EnvironmentSatisfaction', 'Gender',
       'HourlyRate', 'JobInvolvement', 'JobLevel', 'JobRole',
       'JobSatisfaction', 'MaritalStatus', 'NumCompaniesWorked', 'OverTime',
       'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfaction',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole',
       'YearsSinceLastPromotion', 'YearsWithCurrManager'],
      dtype='object')
```

Figure 5: Feature Selection on Attrition

Below are the metrics after predicting **Job Satisfaction** as target column across various machine learning techniques, after applying Smote technique:

Model	Accuracy	F1-Score
Random Forest	0.43	0.43
Logistic Regression	0.30	0.30
AdaBoost	0.33	0.33
K-NN	0.36	0.33
SVM	0.36	0.36
Naive Bayes	0.32	0.32
MLP	0.31	0.31

Below are the metrics after predicting **Environment Satisfaction** as target column across various machine learning techniques, after applying Smote technique:

Model	Accuracy	F1-Score
Random Forest	0.42	0.42
Logistic Regression	0.31	0.31
AdaBoost	0.33	0.34
K-NN	0.35	0.32
SVM	0.33	0.33
Naive Bayes	0.33	0.32
MLP	0.32	0.31

Below are the metrics after predicting **Attrition** as target column across various machine learning techniques, after applying Smote technique:

Model	Accuracy	F1-Score
Random Forest	0.91	0.91
Logistic Regression	0.79	0.79
AdaBoost	0.86	0.86
K-NN	0.84	0.86
SVM	0.87	0.87
Naive Bayes	0.70	0.70
MLP	0.82	0.82

Below is the ROC curve depicting performance of various models for predicting attrition:

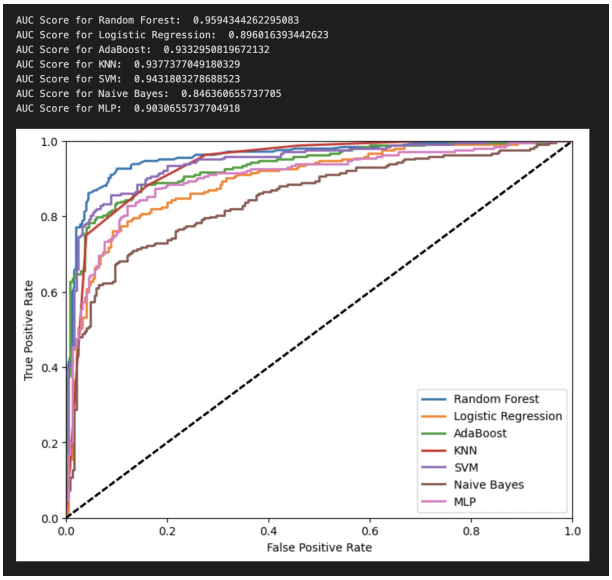

Figure 6: ROC Curve - Attrition

Next, when we predict Job and Environment Satisfaction, and utilize that to predict attrition we see the following metrics:

Model	Accuracy	F1-Score
Random Forest	0.88	0.88
Logistic Regression	0.65	0.65
AdaBoost	0.83	0.83
K-NN	0.63	0.62
SVM	0.61	0.61
Naive Bayes	0.70	0.70
MLP	0.67	0.66

Ensemble ML Models

An Ensemble of following 7 models is created and run for 11 iterations with custom Boosting logic:

(1) Random forest
(2) Logistic regression
(3) Adaboost
(4) KNN
(5) SVM
(6) MLP
(7) Decision Tree

We have used the above seven different machine learning algorithms as the base models for our ensemble. These algorithms may have different strengths and weaknesses, but by combining them together, we can create a more robust and accurate ensemble.

We have used the hard voting method for each iteration of the ensemble. In each iteration, we will train each of the seven base models on a different subset of the 50% of training data, using a different random seed to ensure that each model is different. These random samples are selected based on weights, which we are changing for each iteration by averaging all updated weights by all base models by calculating alpha values similar to the AdaBoost algorithm. Then, we will use each of these models to make predictions on the test data and take the majority vote of the predictions as our final prediction for that iteration.

After all iterations are complete, we will have a set of models from each individual iteration in the ensemble. We then used the hard voting method again to combine these predictions of each iteration into a single final prediction. The final prediction will be the majority vote of the predictions from all of the individual models and individual iterations in the ensemble.

Below find the output of last three iterations:

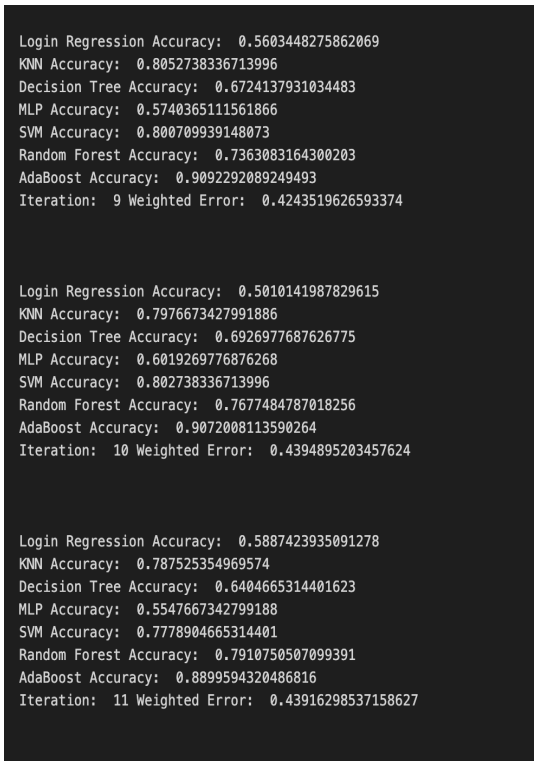

Figure 7: Iteration 9-11

Finally, the confusion matrix report of this Ensemble technique is:

As we can see the overall accuracy comes out to be 0.85 for the trained model.

```
Accuracy Score:  0.8461538461538461
Confusion Matrix:
 [[215  30]
 [ 46 203]]
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.88      0.85       245
           1       0.87      0.82      0.84       249

    accuracy                           0.85       494
   macro avg       0.85      0.85      0.85       494
weighted avg       0.85      0.85      0.85       494
```

Figure 8: Confusion Matrix

Clustering

Clustering with kmeans has been implemented after eliminating unnecessary features. The unnecessary features were decided based on two factors, firstly, the ones which won't be available for new employees like TrainingTimeLastYear, OverTime, etc. and the features which don't really impact the prediction, like EmployeeId, EmployeeCount, etc.

```
Data columns (total 16 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   BusinessTravel            1470 non-null   object
 2   Department                1470 non-null   object
 3   DistanceFromHome          1470 non-null   int64
 4   Education                 1470 non-null   int64
 5   EducationField            1470 non-null   object
 6   Gender                    1470 non-null   object
 7   JobLevel                  1470 non-null   int64
 8   JobRole                   1470 non-null   object
 9   MaritalStatus             1470 non-null   object
 10  MonthlyIncome             1470 non-null   int64
 11  NumCompaniesWorked        1470 non-null   int64
 12  RelationshipSatisfaction  1470 non-null   int64
 13  StockOptionLevel          1470 non-null   int64
 14  TotalWorkingYears         1470 non-null   int64
 15  YearsInCurrentRole        1470 non-null   int64
```

Figure 9: Shortlisted Features - Clustering

Next, number of clusters were determined utilizing elbow method. In our case as we see in the image below, the ideal number of clusters came out to be 4 as that is when we see a spike in the graph.

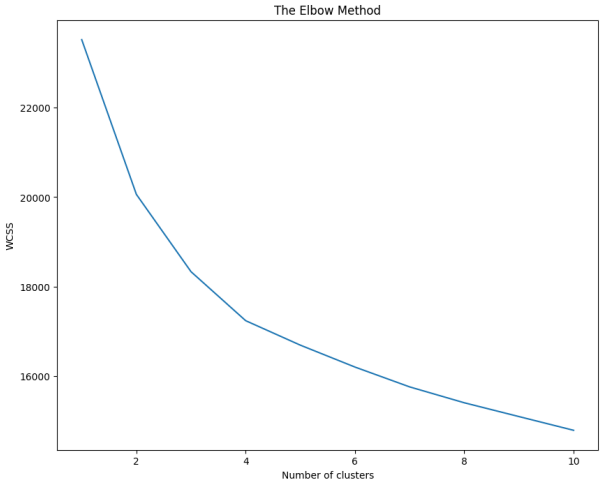

Figure 10: Elbow Curve

Next, we applied clustering and saw the following summary of results:

Cluster	Age	Distance	Income	Years	Attrition
1	47.3	9.0	15141.9	24.6	Y(19),N(230)
2	32.7	8.5	3912.7	7.5	Y(93),N(320)
3	36.9	9.9	4695.5	9.5	Y(41),N(385)
4	34.7	9.3	5687.8	8.8	Y(84),N(298)

Identifying Bias

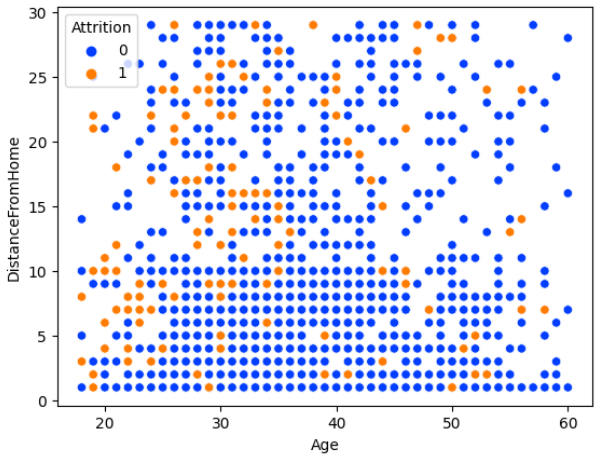

Figure 11: DistanceFromHome vs Age - Attrition

In the above graph of DistanceFromHome vs Age, for predicting attrition, where 1 is Yes (Orange) and 0 is No (Blue), we see that more orange dots are crowded towards younger people and higher distance from home.

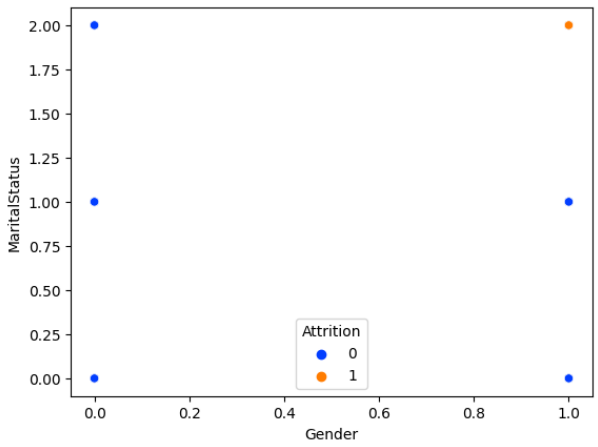

Figure 12: Marital Status vs Gender - Attrition

In the above graph of Marital Status vs Gender, for predicting attrition, where 1 is Yes (Orange) and 0 is No (Blue) for attrition, and for Gender 0 is Female, 1 is Male, and for Marital Status 0 is Single, 1 is Married and 2 is divorced. We observe that the data predicts Divorced Males with attrition yes.

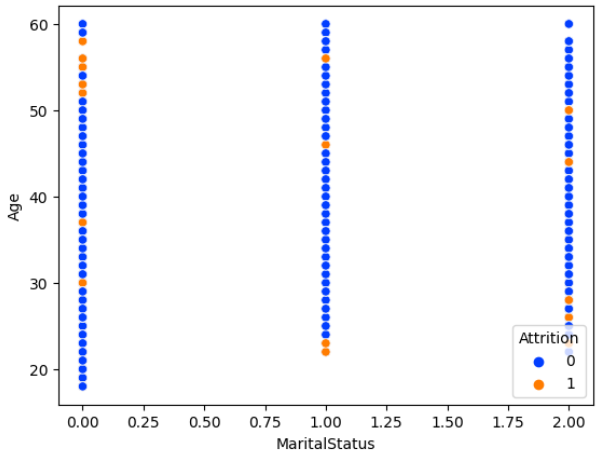

Figure 13: Marital Status vs Age - Attrition

In the above graph of Marital Status vs Age, for predicting attrition, we observe that the data predicts the following combinations with Attrition yes, $\{Older, Single\}$, $\{Younger, Divorced\}$

Removing Biased Features

We now remove the following biased features and retrain the Ensemble of models:

(1) Gender
(2) Age
(3) Marital Status
(4) Distance From Home

Below is the confusion matrix report after retraining and evaluating the Ensemble of models without biased features:

```
Accuracy Score:  0.8340080971659919
Confusion Matrix:
 [[214  31]
 [ 51 198]]
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.87      0.84       245
           1       0.86      0.80      0.83       249

    accuracy                           0.83       494
   macro avg       0.84      0.83      0.83       494
weighted avg       0.84      0.83      0.83       494
```

Figure 14: Confusion Matrix - Unbiased Model

The updated accuracy comes out to be 0.83

Discussion

The graph 3 depicts that while Environment Satisfaction and Job Satisfaction have almost equal representation of all classes, it seems like attrition has much fewer data representing class 1 as compared to class 0. Thus applying Smote technique to evenly distribute these classes would help in increasing accuracy and remove the bias toward a particular class.

The image 5 depicts the shortlisted features that were used in predicting attrition. This does contain potential bias features which we will later analyze whether to keep or remove based on our analyses.

The above results of comparison of various models indicate that environment satisfaction and job satisfaction have overall low accuracy and thus aren't that effective when predicted independently.

In the case of attrition, Random Forest, SVM, and KNN are the top three when considering the evaluation metrics. Thus these can be potential candidates for using these to create an Ensemble model.

We made an interesting observation here, we had initially considered that using the predicted values of job satisfaction and environment satisfaction as input features for attrition prediction would increase the model's accuracy, which didn't work in our favor. Turns out that due to the less independent accuracy of these two attributes, even when combined with attrition, they prove to be disadvantageous for attrition prediction. The tables above clearly depict this claim, we see that for all the models except for Naive Bayes, the accuracy and F1-score decreased, thus indicating that the model didn't perform that well.

Utilizing seven different ML techniques to predict output and applying boosting to penalize the incorrect predictions and reward the correct ones, helped create a pretty accurate model. We got an

accuracy of 0.85 as we can see in figure 8 and considering it's from across seven different techniques, the chance of model overfitting to this particular dataset reduces considerably.

In the case of clustering, firstly 4 clusters were created based on the elbow method interpretations from the figure 10. We noticed that K-Means created a cluster 4 containing only the Sales department, so whenever any sales employee was predicted, there was a high chance it would automatically be assigned to cluster 4. Apart from that, the major point of the division was Age and years of experience as we see in table 4.

In terms of analysis and observations, we get some pretty interesting results. From figure 11, we see that there is a high rate of attrition among younger people as compared to older people. From figure 12 we observe that current data shows that being a divorced male, there are higher chances of attrition. Finally, in figure 13, we observe that the following combinations $\{Older, Single\}$, $\{Younger, Divorced\}$ are predicted to have a higher chance of attrition.

The above observations clearly show that if these features are kept in the dataset, the model will be biased towards a particular section of society and this must not be the case. There are chances that the model will not select qualifying young people only because of their age, or it might not select Divorced people because of their marital status. Thus these biases must be eliminated from the dataset and the model must be retrained.

Figure 14 shows that after removing biased features, the accuracy of the model reduces and this was expected. It is better to have a bit less accurate model than to have biased features in the model.

Source Code

The source code for this project can be found in the following GitHub Repository: (https://github.com/deep-mm/Potential-Performance-Predictor).

5 CONCLUSION

The ensemble of machine learning models with boosting was really effective and gave promising accuracy. This novel aspect of the project was inspired from Adaboost. Utilizing multiple machine learning techniques would help avoid overfitting and will work effectively on new data as well.

Identifying bias was a crucial part of this project from a ethical point of view. Whenever the decisions of model impacts human lives, bias should never be ignored. We saw how the features age, gender, marital status and distance from home had a possibility of bias against a particular section of people in society. Thus, we removed these features, even though it meant a reduction in accuracy of our models.

6 MEETING ATTENDANCE

A total of 8 meetings were held in the past month among team members (All team members were present during these meetings):

(1) March 8th, 2023
(2) March 22nd, 2023
(3) March 27th, 2023
(4) March 31st, 2023
(5) April 5th, 2023
(6) April 12th, 2023
(7) April 19th, 2023
(8) April 23rd, 2023

REFERENCES

[1] Chen-L. Chein, C. 2006. *Data mining to improve personnel selection and enhance human capital: A case study in high technology industry.* Expert Systems with Applications, In Press.
[2] E. Kayha. 2007. *The Effects of Job Characteristics and Working Conditions on Job Performance.* International Journal of Industrial Ergonomics, In Press.
[3] P. Kotsiantis, Sotiris & Pintelas. 2005. *Combining Bagging and Boosting.* International Journal of Computational Intelligence.
[4] Sajedi Hosseini F. Choubin B. et al. Mosavi, A. 2021. *Ensemble Boosting and Bagging Based Machine Learning Models for Groundwater Potential Prediction.*
[5] Eman Al Nagi Qasem A. Al-Radaideh. 2012. *Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance.* International Journal of Advanced Computer Science and Applications.