Absenteeism at Work Prediction

Ashwin Rajh

Table Of Contents

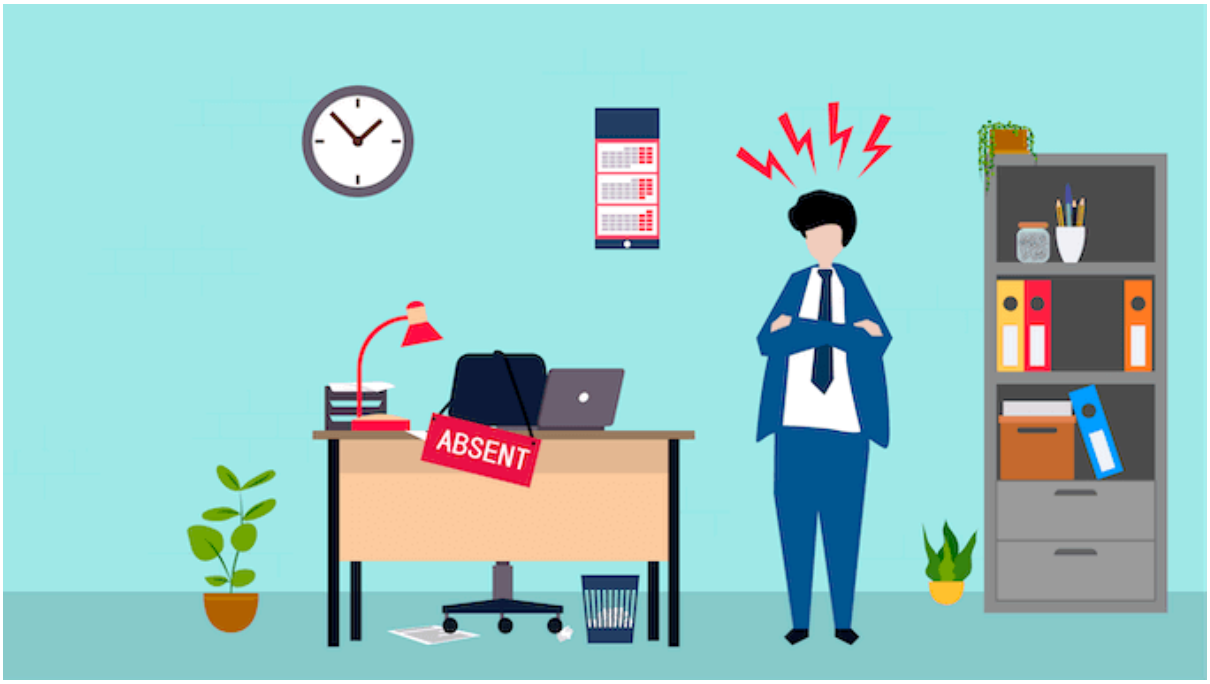
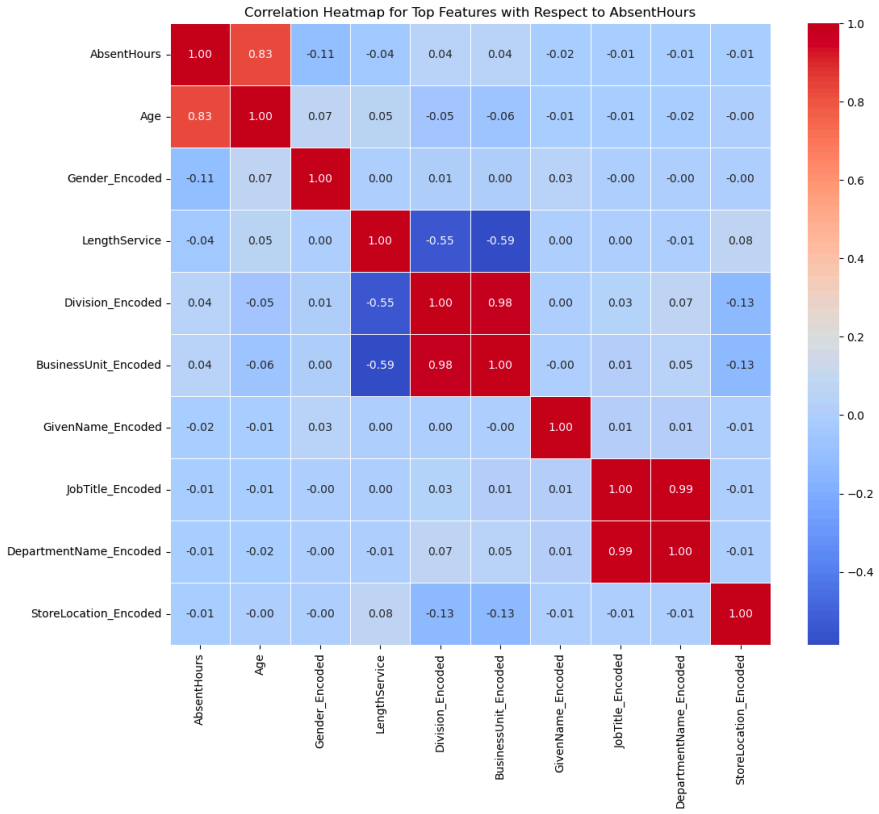
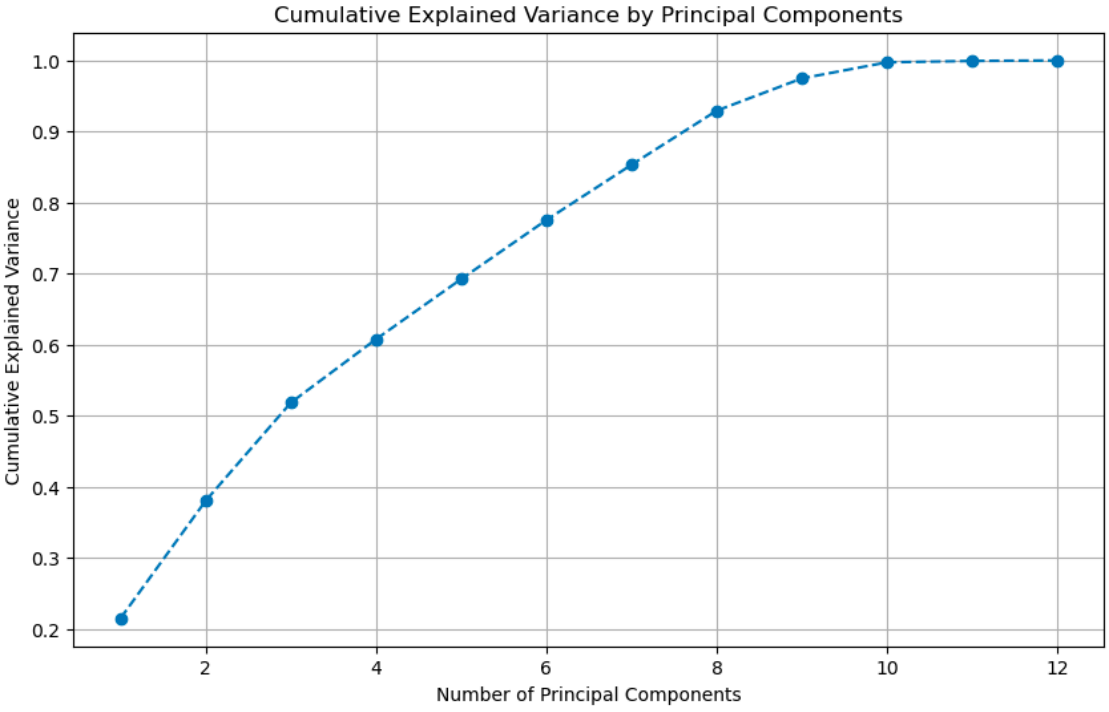


Table Of Charts:

Correlation Heatmap for Top Features with Respect to AbsentHours:



Cumulative Explained Variance by Principal Components:



Abstract:

This project delves into predicting absenteeism at the workplace, utilising an extensive HR dataset. The dataset encompasses various employee details, including identification numbers, names, gender, city, job title, department, store location, business unit, division, age, length of service, and a pivotal variable—the number of hours absent. Our objective is to develop machine learning models for forecasting the 'No. of hours absent.'

The emphasis is on minimising Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) while maximising the R-Square score. We employ Multiple Linear Regressor (MLR) with Ridge Regression, Random Forest Regressor (RFR), MLR with Principal Component Analysis (PCA), and RFR with PCA to achieve these goals.

Through this project, we explore and understand the realm of HR analytics, gaining insights into absenteeism patterns and bolstering our ability to effectively manage workforce dynamics.

Introduction:

Absenteeism poses multifaceted challenges to organizations, including disruptions in workflow, increased workload on remaining staff, and potential financial implications. Traditional approaches to managing absenteeism may lack precision and fail to address underlying causes, leading to suboptimal outcomes. Consequently, there is a pressing need for advanced predictive models that leverage data-driven insights to forecast absenteeism accurately. By understanding the historical patterns and factors contributing to absenteeism, companies can implement targeted strategies to enhance employee well-being and maintain a resilient and productive workforce.This project aims to tackle these challenges head-on by developing robust machine learning models capable of predicting the number of hours an employee is likely to be absent.This project aims to tackle these challenges head-on by developing robust machine learning models capable of predicting the number of hours an employee is likely to be absent

Software and Libraries Used:

Pandas, Numpy, Matplotlib are used for data manipulations and data visualisations.
Scikit-Learn is utilized for preprocessing, implementing and evaluating the performance of models such as Logistic Regression(Ridge Regression) and RandomForestRegressor with and without PCA.

```
from sklearn.preprocessing import LabelEncoder
import pandas as pd
import numpy as np
from collections import Counter
from sklearn.linear_model import Ridge
from sklearn.ensemble import RandomForestRegressor
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import KFold, cross_val_predict
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
import matplotlib.pyplot as plt
from sklearn.metrics import accuracy_score, precision_score, recall_score
```

EDA:

LabelEncoder was used to convert features of type object to type int.Used Standardisation on X(features) to ensure no discrepancies. Took K=10 in terms of features to find correlation heat map wrt AbsentHours. Standardised all the features and kept ready to be used for model.

Algorithm:

Set the Ridge Regression model with a specific regularization parameter (alpha=0.1). Define the number of folds for cross-validation (k_folds = 5). Create a KFold cross-validation object (kf) with 5 splits, enabling shuffling of data, and setting a random seed for reproducibility. Initialize empty lists (mse_scores, rmse_scores, r2_scores) to store performance metrics for each fold. Iterate through each fold using the KFold object (kf.split(X_scaled)).Split the dataset into training and testing sets based on the current fold indices.Train the Ridge Regression model on the training set (ridge_model.fit(X_train, y_train)).Predict the target variable (y_pred) on the test set.Calculate Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Squared (r2) scores.Append the calculated scores to their respective lists.bAfter all folds have been processed, find the minimum MSE, minimum RMSE, and maximum R-Squared score among

the collected scores. Display the minimum MSE, minimum RMSE, and maximum R-Squared score.

Algorithm is similar for RandomForestRegressor except here the model becomes rfr_model = RandomForestRegressor(random_state=1).

For PCA:

Standardize X and apply a PCA object to it using pca.fit_transform(X_standardized). Obtain the explained variance ratio for each principal component using pca.explained_variance_ratio(). Plot the cumulative explained variance against the number of principal components. X-axis: Number of Principal Components, Y-axis: Cumulative Explained Variance. Display the plot with markers, lines, and appropriate labels. It helps in visualising the the cumulative explained variance to aid in determining the optimal number of principal components. Calculate the cumulative variance explained by each principal component using np.cumsum(explained_variance_ratio). Determine the index of the first cumulative variance value greater than or equal to 95% using np.argmax(cumulative_variance >= 0.95). Add 1 to the index to obtain the number of components required for retaining at least 95% of the variance. Use array slicing to select the principal components up to the identified number for retaining 95% variance. The resulting array X_selected_pca contains the selected principal components, retaining at least 95% of the variance. The selected principal components can then be used for further analysis or model training. Using the X_selected_pca array, we can further train data for both Ridge and Random Forest Regressor.

Evaluation Metrics:

*Mean Squared Error***:** MSE measures the average of the squared differences between predicted and actual values.

*Root Mean Squared Error***:** RMSE is the square root of the MSE, providing a measure of the average magnitude of errors.

R-Square Score: R-Square, or the coefficient of determination, measures the proportion of variance in the dependent variable explained by the independent variables.

Ridge Regression:

```
#Ridge Regression
```

```
Minimum Mean Squared Error (MSE): 593.5243041638419
Minimum Root Mean Squared Error (RMSE): 24.362354240997355
Maximum R-Square Score: 0.7513197417783806
```

RandomForestRegressor:

```
#Random Forest Regression model
```

```
Minimum Mean Squared Error (MSE): 558.6883941154392
Minimum Root Mean Squared Error (RMSE): 23.636590154153776
Maximum R-Square Score: 0.7659156109710101
```

Now using Principal Component Analysis(PCA):

Ridge Regression:

```
#PCA Framework - ridge regression model
```

```
Minimum Mean Squared Error (MSE): 593.0751871479505
Minimum Root Mean Squared Error (RMSE): 24.353135057892455
Maximum R-Square Score: 0.7515079169461034
```

Random Forest Regression:

```
#PCA Framework - RandomForest regression model
```

```
Minimum Mean Squared Error (MSE): 568.459838192318
Minimum Root Mean Squared Error (RMSE): 23.842395814857156
Maximum R-Square Score: 0.7618214816839884
```

Result:

The RandomForest model without PCA emerged as the top-performing model, showcasing superior predictive capabilities across all metrics. Therefore, the most effective in predicting absenteeism, exhibiting lower MSE and RMSE alongside a higher R-Square score.

Conclusion:

The ultimate goal was to construct machine learning models that forecast the number of hours absent, with a focus on minimizing Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) while maximizing the R-Square score. We explored the efficacy of Multiple Linear Regressor (MLR) with Ridge Regression, Random Forest Regressor (RFR), MLR with Principal Component Analysis (PCA), and RFR with PCA.

Through rigorous experimentation and analysis, the RandomForest regression model without PCA emerged as the standout performer, showcasing superior predictive capabilities. This model demonstrated lower MSE and RMSE, coupled with a higher R-Square score, making it the preferred choice for accurately forecasting absenteeism.