Complete Productivity Analysis of Employees with Machine Learning

AIN 429

EFE EMİRHAN DOĞAN

2230765023

1. Abstract

This project investigates predicting productivity score of employees using machine learning algorithms. A dataset that contains 30 attributes without missing value like ‘Years_of_Experience’, ‘Manager_Support_Level’, ‘Job_Satisfaction’… were used. Barplots, Histplots, Scatterplots and Boxplots were used for visualizing and understand dataset. Different encoding techniques were also used to encoding categorical variables to feed regression models. Multiple regression and classification algorithms were used, including Linear Regression, Random Forest Regressor, Support Vector Regressor for regression and Logistic Regression, Random Forest Classifier and Voting Classifier for classification. The regression model achieved 0.92 score of R^2 which proves strong predictive performance. And The classification model achieved 0.89 score of accuracy.

2. Introduction

Worker productivity is one of the most debated topics in the modern world. Survey data collected from employees was used in this project. An attempt was made to found which factors affect worker productivity. Regression models were used to predict productivity scores. Subsequently, the productivity scores were divided into classes and classification models were used.

Data visualisations and analyses were performed on the data using the pandas library. The relations between Productivity and Age, Experience, Menager Support Level were investigated. But because of high dimensionality of data and possible complex relations these analyses were not enough. At this point machine learning algorithms are required.

Regression models were used to predict value of ‘Productivity_Score’. And different evaluation metrics were used to estimate model’s performances and compare them. These metrics are Mean-Squered Error and R^2. MSE and R^2 might be meaningless on their own so they were used together. The smaller the MSE value the better it is. If $R^2 < 0$, it means model worse than average. The 0.3-0.6 value of R^2 indicates average and greater values means good model. But it should be checked to avoid overfitting. CV (Cross-Validation) Score was used to estimate is model overfitted or not.

Cross Validation is a method that shows how machine learning model performs on unseen data. It splits data into parts, trains the model with some parts and test the model with others. Repeats this process by choosing different parts of dataset every time.

3. Methodology

3.1 Data Preprocessing and Insights

Dataset contains 30 attributes and 1500 non-null samples. ‘Employee_ID’ and ‘Survey_Date’ have been dropped due to being unnecessary. Correlation heatmap for numerical columns was plotted to explore correlation between them. It clearly shows that there are some features correlated with ‘Production_Score’. They’re: ‘Task_Completion_Rate’, ‘Quality_Score’, ‘Innovation_Score’, ‘Efficiency_Rating’ and ‘Job_Satisfaction’. Also, boxenplot shows that ‘Manager_Support_Level’ affects our target. This might very important observation.

The average productivity score for employees with over 10 years of experience is 88.5, while the average productivity score for employees with less than 10 years of experience is 82.4. The average productivity score for employees who are older than 40 is 87, while the average productivity score for employees who are younger than 40 is 81.7. These statistics clearly illustrate the relationships between experience, age and productivity.

Also according to answers if employees have poor home office quality, their average productivity score is 76.8, if they have average home office quality, average productivity score is 85.4, if they have good home office quality, average productivity score is 86.4 and if they have excellent home office quality, average productivity score is 91.2. These statistics clearly illustrate that having better quality of home office is affects productivity positively.

To encoding categorical variables, one hot encoding has been used for nominal variables and ordinal encoding for ordinal variables. Then their types were converted Boolean to float for consistency.

Z-Score standardization was used to scale data before feed the models. Z-Score standardization sets their mean to ‘0’ and sets their standard deviation to ‘1’. It is important especially for some algorithms like Linear Regression, Logistic Regression, SVM.

Dataset was split into train and test by using scikit-learn library. Test size has given as 0.2. And final size for train was (1200,58) and for test (300,58).

3.2 Modeling

Linear Regression, Random Forest Regressor and Support Vector Classifier were used in order. Linear Regression is one of the simplest regression algorithms and others more complex then Linear Regression. All models were evaluated by MSE, RMSE and R^2. Following by cross validation score to get more strong evaluation results. The feature importance were extracted by using method of Random Forest Regressor. According to it most three important features are: ‘Efficiency_Rating’, ‘Quality_Score’, ‘Task_Completion_Rate’.

To show how different values of specific features affect prediction. Min and max values of ‘Job_Satisfaction’, ‘Manager_Support_Level’ and ‘Age’ have given to models and all other values were kept same. As result it showed that ‘Job_Satisfaction’ alone can affect prediction.

For classification part, target attribute has divided into 3 parts. The target feature ‘Productivity_Score’ has minimum value -3.2 and max value 0.98. So it has divided as: if (y <= -2) -> 0 (Low), elif (y<0) -> 1 (Average), else ->2 (High). Then Logistic Regression, Random Forest Classifier, Decision Tree Classifier and Voting Classifier have applied in order. All models have evaluated by classification report (f1 score, precision, recall, accuracy) and plotted confusion matrixes.

4. Results

4.1 Regression Results

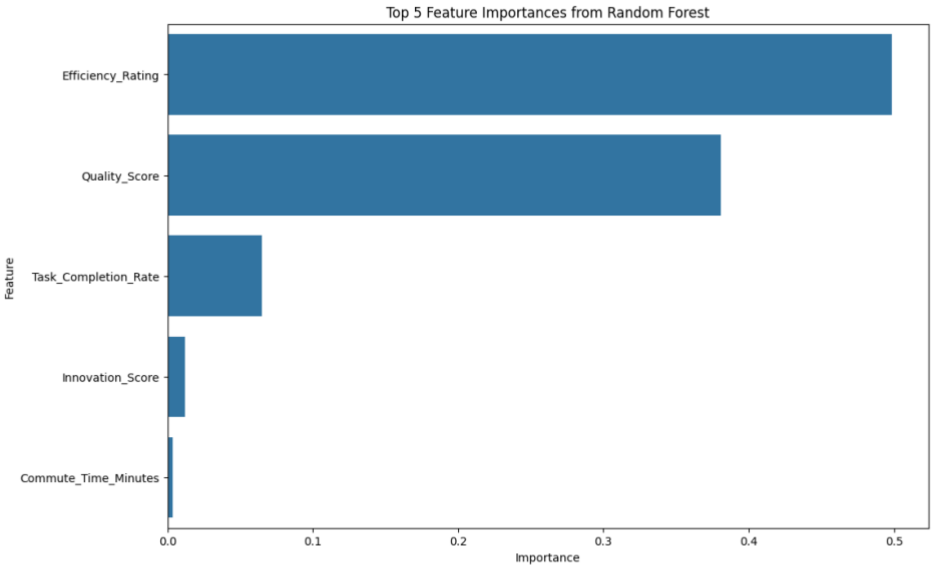
Linear Regression: RMSE = 0.26, $R^2 = 0.92$, CV $R^2 = 0.92$

Random Forest: RMSE = 0.26, $R^2 = 0.92$

Support Vector: Regressor RMSE = 0.30, $R^2 = 0.90$

It can be said that regression models performed quite robustly and accurately according to the results.

Feature importances have extracted by using method of Random Forest Regressor



For what if analysis for Job_Satisfaction, Manager_Support_Level and Age attributes minimum and maximum values have tested.

Job_Satisfaction: Min = 37.6, Predicted Productivity = 79.7 – Max = 100, Predicted Productivity = 83.2

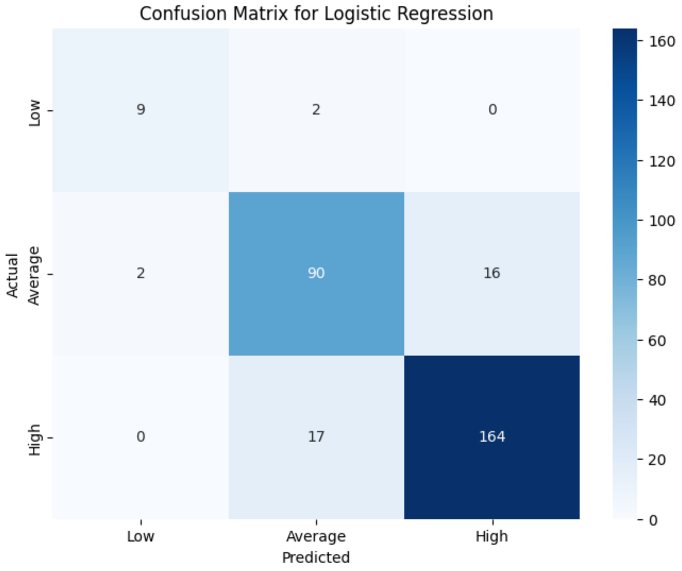
Manager_Support_Level: Min = 0, Predicted Productivity = 82.9, Max = 4, Predicted Productivity = 82.9

Age: Min = 22, Predicted Productivity = 82.4, Max = 65, Predicted Productivity = 84

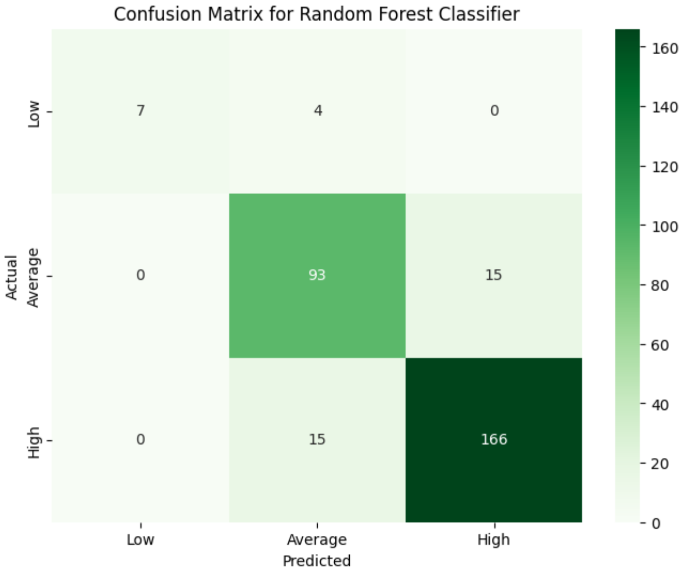
It can be said Manager_Support_Level there is no effect on productivity only by itself unlike Job_Satisfaction and Age according to the results.

4.2 Classification Results

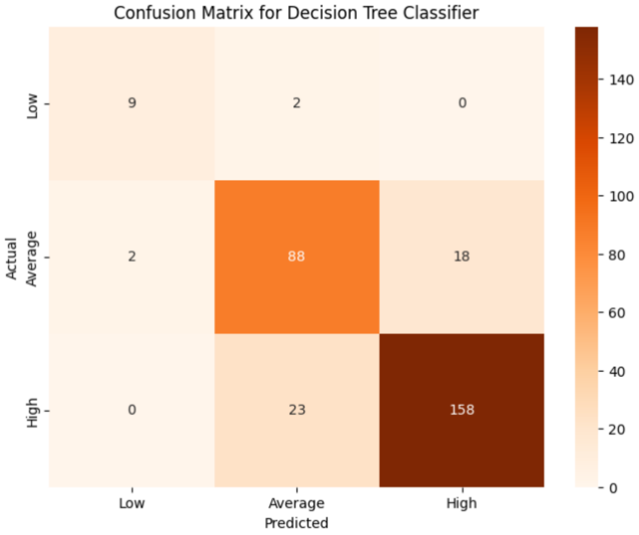
Logistic Regression: Accuracy = 0.88, Macro F1 = 0.85, Weighted F1 = 0.88



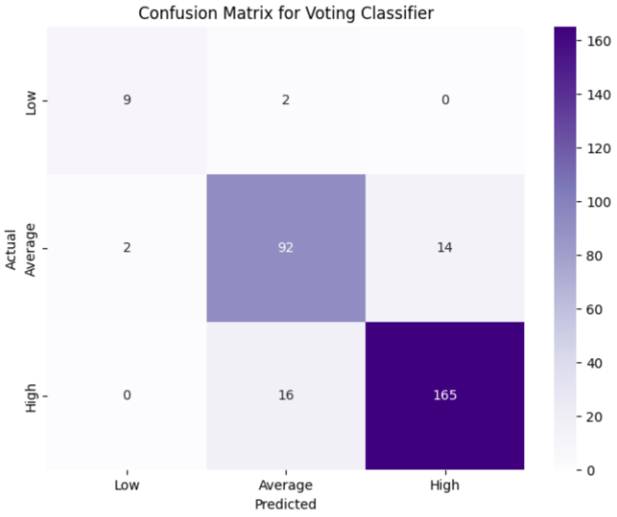
Random Forest Classifier: Accuracy = 0.89, Macro F1 = 0.85, Weighted F1 = 0.89



Decision Tree Classifier: Accuracy = 0.85, Macro F1 = 0.83, Weighted F1 = 0.85



Voting Classifier: Accuracy = 0.89, Macro F1 = 0.86, Weighted F1 = 0.89 (Logistic Regression, Random Forest, Decisiton Tree were used, voting = soft to predicts the class label based on the argmax of the sums of the predicted probabilities.)



5. Discussion/Conclusion

This study analysed the factors affecting employee productivity. The results show that although there are several features that significantly affect the productivity score, the productivity score is influenced by many factors. And again, according to the results, it can be said that a complex model does not necessarily mean a better model. Employers can use this type of analysis to plan what needs to be done to increase employee productivity.