Report

Name: Upesh Jeengar
Mail: upeshjeengar@gmail.com

To view in notebook form:
https://colab.research.google.com/drive/1OqtmjbeJdV9Taezv0iqY97u_q7s7EIxW?usp=sharing

1.Dataset Analysis and Preprocessing

Dataset https://raw.githubusercontent.com/Upeshjeengar/IBM-HR-Analytics-Employee-Attrition-Performance/main/data.csv

(Downloaded from Kaggle)

Dataset has no missing values and outliers

Plotted histograms for all features using matplotlib library

Used Label encoder to transform categorical variables(string type) to integer type

2.Model Development

Tried different possible classification Models to predict Attrition. Here are some results:

Index	Random Forest	Logistic	SVM
Accuracy	0.84	0.83	0.83
Precision	0.6	0.66	0.0
Recall	0.1	0.03	0.0
F1-score	0.17	0.06	0.0

3.Model Evaluation and Optimization:

1. Accuracy: The Random Forest model has the highest accuracy (0.84), followed closely by the Logistic Regression model (0.83), and then the SVM model (0.83).

2. Precision: Precision measures the ratio of correctly predicted positive observations to the total predicted positives. Here, the Logistic Regression model has the highest precision for positive

predictions (0.66), followed by the Random Forest model (0.600000). The SVM model has a precision of 0.00 for positive predictions, which could indicate a problem with classifying positive cases.

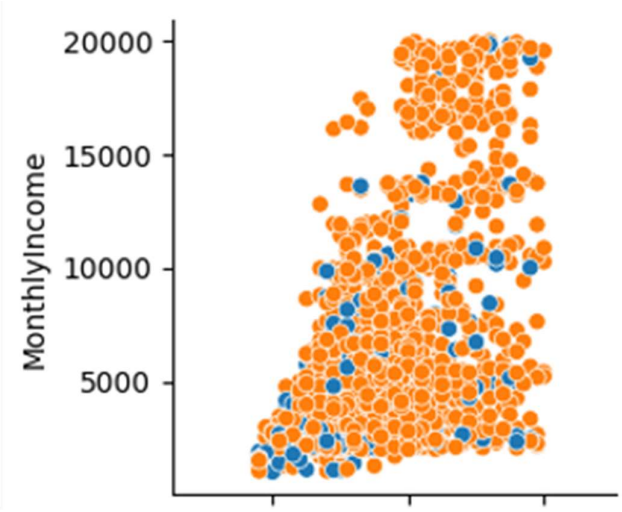
3. Recall: Recall measures the ratio of correctly predicted positive observations to all actual positives. In this case, the Random Forest model has the highest recall (0.10), followed by the Logistic Regression model (0.03), and again, the SVM model has a recall of 0.00 for positive cases.

4. F1-score: The F1-score is the harmonic mean of precision and recall and is a good overall measure of a model's performance. The Random Forest model has the highest F1-score (0.17), followed by the Logistic Regression model (0.06), and the SVM model has an F1-score of 0.00 for positive cases.

Based on these scores, the Random Forest model appears to be the best choice among the three models, as it has the highest accuracy, precision, recall, and F1-score. The Logistic Regression model also performs reasonably well but has lower scores compared to the Random Forest model. The SVM model seems to perform poorly in terms of precision, recall, and F1-score, especially for positive predictions.

Factors on which Attrition depends:

1.Low monthly income



2.Bad Job Environment