Playground Series, Season 3 Episode 3- Employee Attrition dataset

Tabular Classification with an Employee Attrition Dataset

Prepared by: Soumee Ghosh (emp id: 2211444)

Data Description:

The dataset for this competition (both train and test) was generated from a deep learning model trained on a Employee Attrition. Feature distributions are close to, but not exactly the same, as the original. Feel free to use the original dataset as part of this competition, both to explore differences as well as to see whether incorporating the original in training improves model performance.

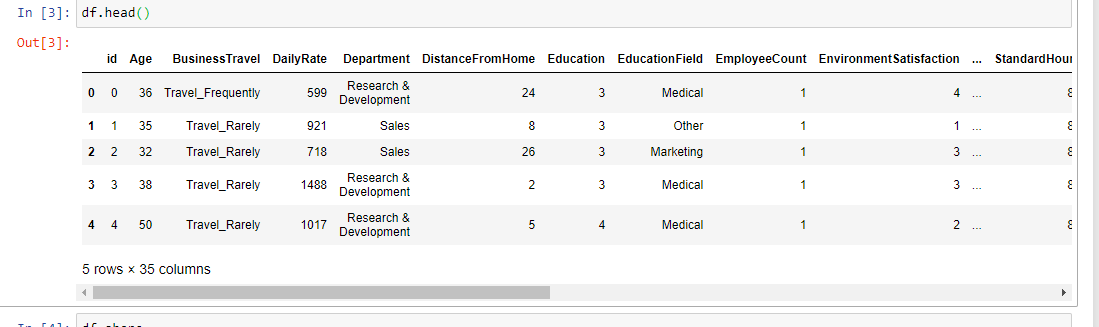
This dataset is a tabular classification dataset about employee attrition which has been taken from Kaggle. The main objective of this problem is to classify whether an employee will leave the job or not based on the features available.

The analysis can be divided into the following steps:

* Understanding the data
* Statistical analysis
* EDA
* Data pre-processing
* Model Building
* Model Evaluation

1. Understanding the data:

The first thing that one should do after loading a dataset is to check the head of the data frame.

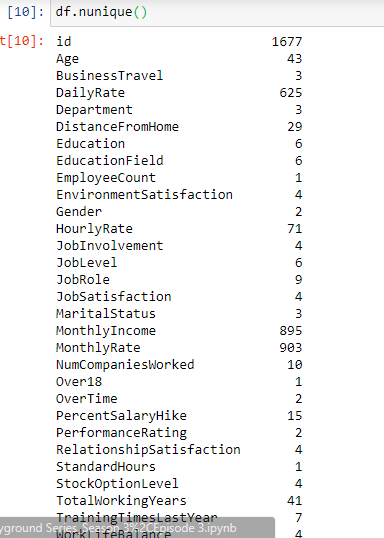


On checking the shape of the dataset, we can observe that there are 1676 rows and 35 columns.

Using the .info() method , we try to see the datatype of the features in the data frame.

Thereafter we perform some statistical analysis and try to look at the distribution of the data.

We try to look for unique values of the features in the dataset.



1. EDA:

Now we perform the exploratory data analysis. Firstly we try to find no of unique categories and their value counts for each feature in the dataset.

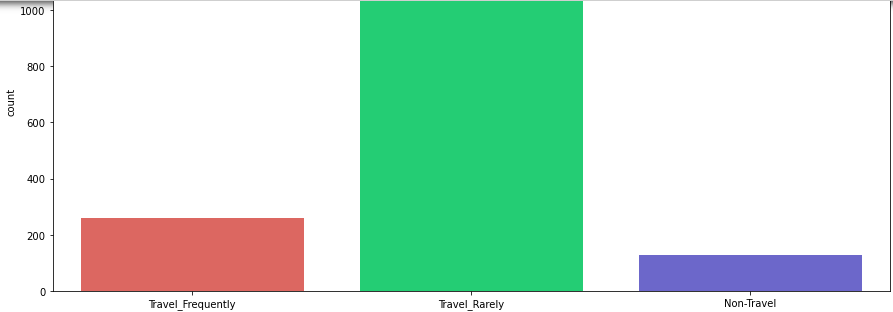
For example for the feature Business Travel we observe that there are three unique categories :

Travel rarely, Travel Frequently, Non-travel and we obtain the necessary value counts i.e. the number of people who opt for each of these.

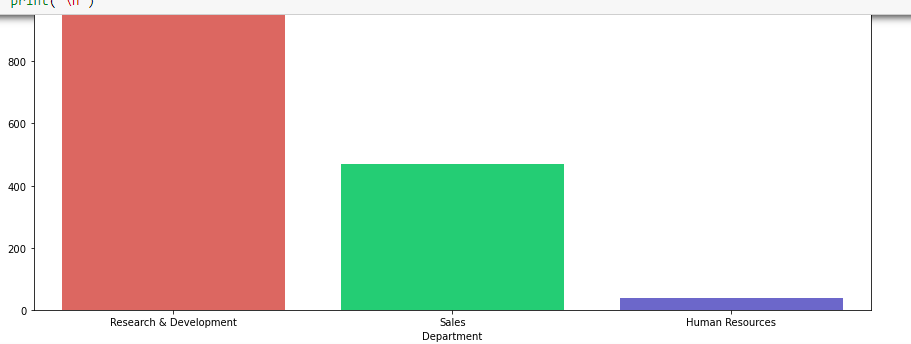
Similarly we try to observe this for other categorical variables as well.

Using a for loop we try to create a visualisation with the help of seaborn and by using count plot .

Here we can observe that in case of Business travel , people travelling rarely are more compared to people travelling frequently and non-travelling people.



Similarly we observe that in case of the Department category , there ae more people who belong to Research ad development compared to sales department and human resources.



For education field, we observe that people belonging to life sciences and medical field are far more compared to other fields.

For that target variable : Attrition we observe that the data is quite imbalanced. The number of people who left the job are 200 compared to those still having the job which is 1477.

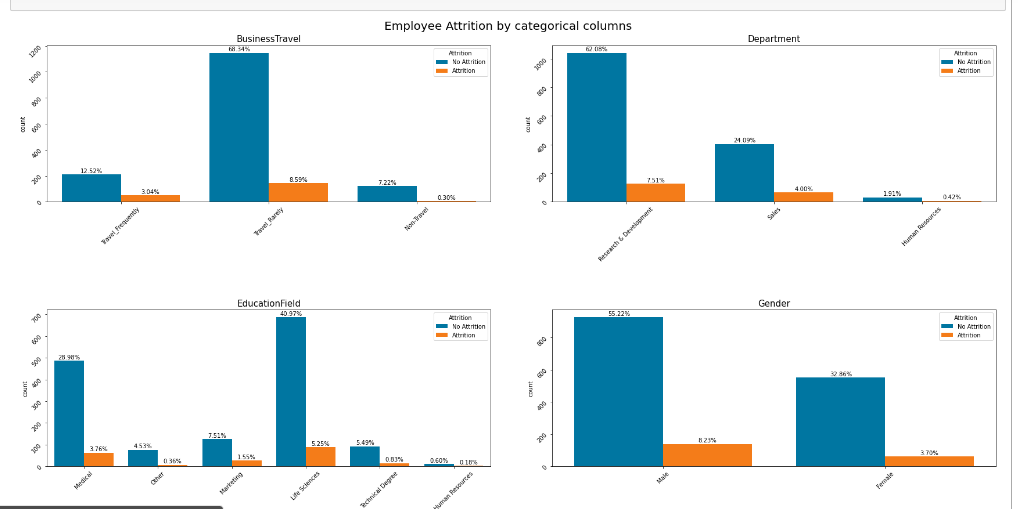
So the dataset is imbalanced.

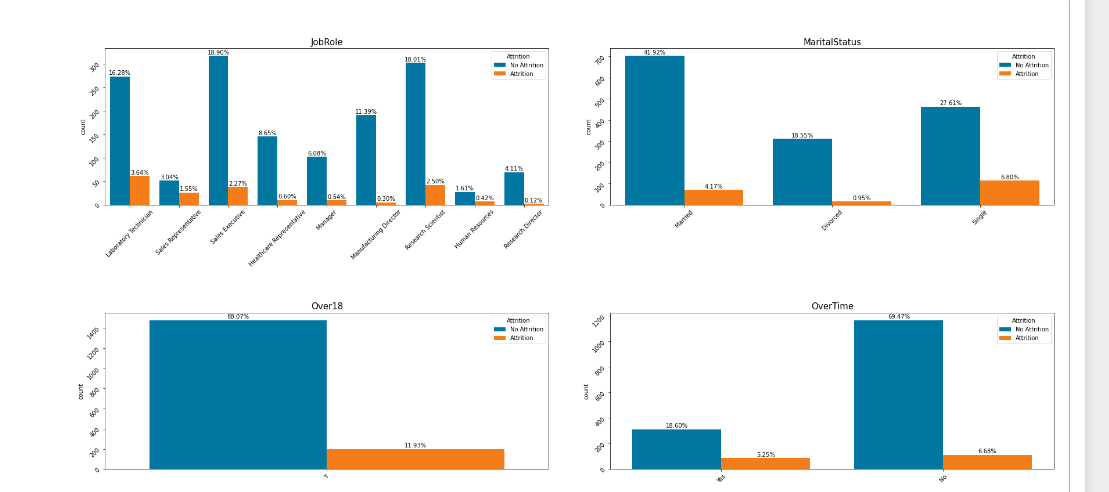
Thereafter we divide the columns into numerical and categorical columns so that we can continue with the univariate and bivariate analysis.

Numerical columns : 'Age', 'DailyRate', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EnvironmentSatisfaction', 'HourlyRate', 'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager', 'Attrition'

Object columns: 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'Over18', 'OverTime'

Thereafter we try to see the employee attrition by categorical columns, we find that attrition is more in people who travel rarely compared to those who travel frequently. We also observe that attrition is more in the research and development business followed by sales department. Attrition is more for males compared to females and it is mostly for people with marital status as single.





Then we check for the distribution of the numerical columns with respect to employee attrition with the help of box plot.

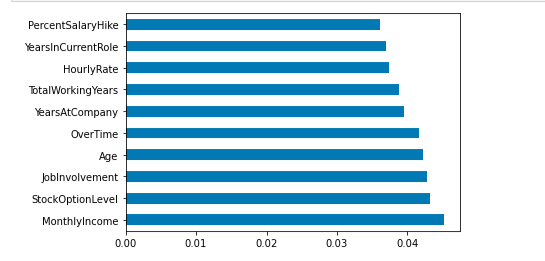
Now before splitting our data and starting the model building we perform encoding on he categorical variables. We use label encoding for this purpose. We check the correlation between the variables to see how much each variable effects the employee attrition.

We then proceed with the data pre processing step. Firstly we use the standard scaler to scale the features.

Since there are quite a number of features so we use ExtratreesClassifier for feature selection.

ExtraTreesClassifier is a machine learning algorithm in the scikit-learn library, which is a variant of decision trees. It is a type of ensemble learning method that combines multiple decision trees and uses them to make predictions. Like other decision tree algorithms, ExtraTreesClassifier works by recursively partitioning the input data into smaller subsets based on the values of its features. However, unlike standard decision trees, which use a single threshold for each feature to determine how to split the data, ExtraTreesClassifier uses random thresholds for each feature. This introduces additional randomness and reduces the variance of the model, which can improve its performance and prevent overfitting.Moreover, ExtraTreesClassifier builds multiple decision trees using random subsets of the training data and random subsets of the features. This technique is called "random subspace method" or "feature bagging". It introduces more diversity among the trees, which can lead to better generalization and robustness. To make a prediction, ExtraTreesClassifier aggregates the predictions of all the trees using either majority voting (for classification) or averaging (for regression).ExtraTreesClassifier is particularly useful when dealing with high-dimensional datasets, noisy data, or imbalanced datasets, where other models might struggle to achieve good performance. It also has relatively low computational requirements, making it scalable to large datasets.

On plotting the feature importance we find the top 10 features.



So we successfully reduced the number of features from 35 to 10 . We divide the data in such a way that these top 10 features are included in X and Attrition is taken as our target variable in y.

We now move to our model building phase.

We split the data into training and validation set and since this is a binary classification problem we perform classification.

The models used are:

* Logistic regression
* Decision Tree Classifier
* Random Forest Classifier

We also performed Hyperparameter Tuning using grid search cv to find the best parameters and best model score.

We finally used the stratified k fold . Stratified k-fold cross-validation is widely used in machine learning as it provides a robust estimate of model performance, especially when the dataset has class imbalance.

Using these we find the optimal models and their best scores upon hyperparameter tuning.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Decision Tree | Random Forest | Logistic Regression |
| Best Score | 0.73 | 0.89 | 0.88 |

We also use deep learning and neural network to train the model using keras sequential layers and obtain a test score of 0.869.

Thus we find that Random forest gives the best result for training our model.