Employee Attrition Predictions

Objectives

Development of Employee Attrition Prediction Model for an Organization which is facing high attrition percentage. This model is determine the decision of an employee if he is going to leave the company or not.

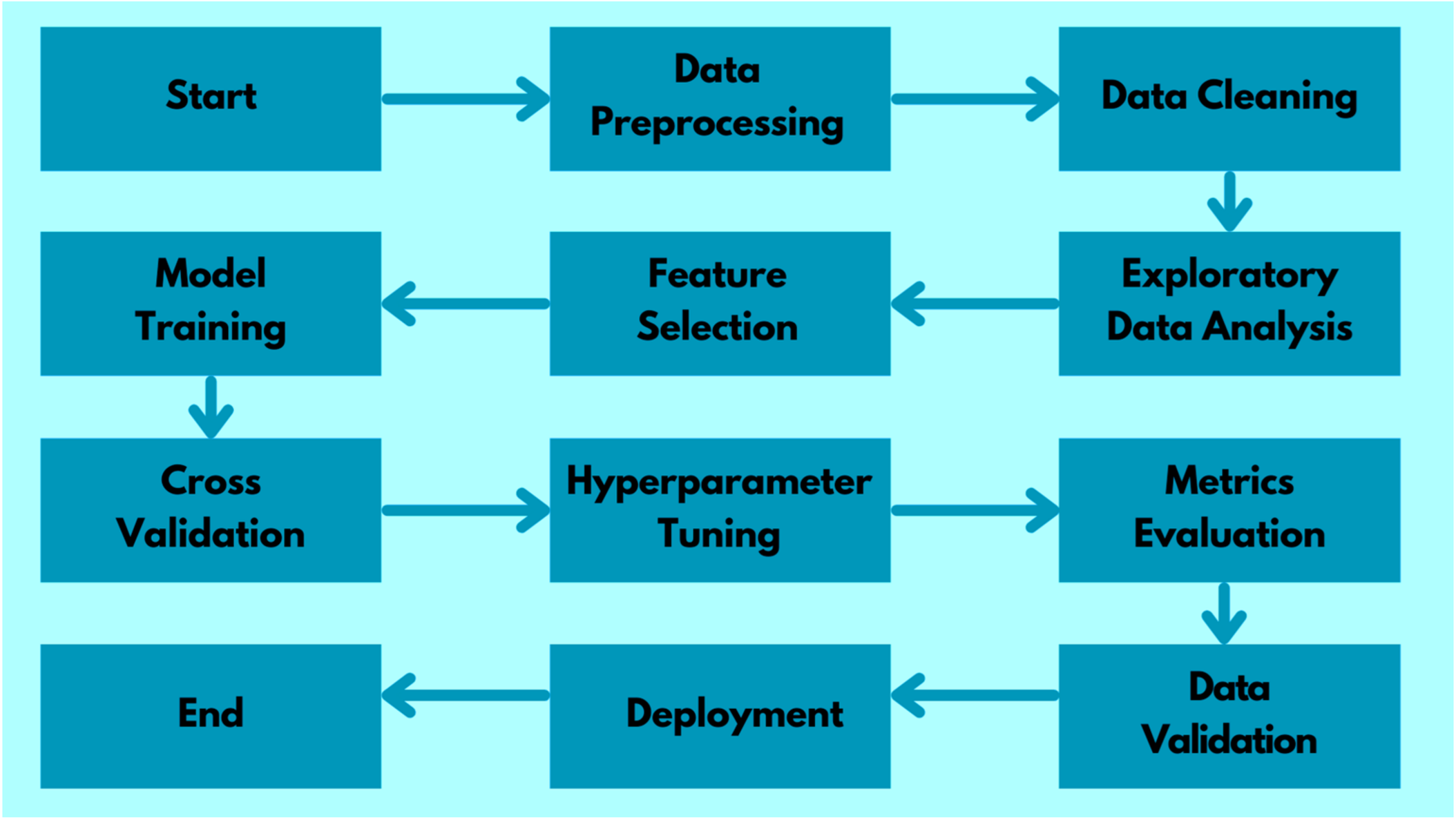
Benefits

- Predict the Employee attrition based on the given parameters
- Provides insights about Employee in the organization.
- Provides the relation between the parameters.
- Provides the decision of employee so that company can take some actions regarding high resignation.

Data Sharing Agreement

- Sample File Name: **WA_Fn-UseC_-HR-Employee-Attrition.csv**
- Cleaned Data File Name: **data.csv**
- Number of Columns: **35**
- Number of Rows: **1470**
- Columns Name: **Age, Attrition, Business Travel, DailyRate, Department, DistanceFromHome, Education, EducationField, EmployeeCount, Employee Number, Environment satisfaction, Gender, HourlyRate, JobInvolvement, JobLevel, JobRole, JobSatisfaction .....(18 more)**
- Columns Datatype: **int64 and object**

Architecture



Step 1: Data Preprocessing and EDA

- Raw data contains 31 columns: Age, Daily Rate, DistanceFromHome, Education, EnvironmentDatisfaction, HourlyRate, JobInvolvement ..etc.
- I reconstruct the dataset by removing some unnecessary data column.
- I removed : EmployeeCount, EmployeeNumber, StandardHours, Over18.
- I also tried to deal with the missing value which is not actually in this dataset.

Step 3: Logistic Regression

- First we split the data into 20% test size and rest 80% to train the model.
- This Model leads us to get accuracy of 88.43% .
- Random Forest when tested on the unseen data, we got pretty near or quite correct values.

Step 4: Random Forest Classifier

- Random Forest Classifier initially we split the data into 20% test size and rest 80% to train the model.
- This Model leads us to get accuracy of 85.71% .
- Random Forest when tested on the unseen data, we got pretty near or quite correct values.

Step 5: Support Vector Machine

- After Random Forest we tried support Vector Machine to check the accuracy is this model is better than the previous used models.
- We split the data into 20% data for the testing and rest 80% for training the model.
- We get 86.73% accuracy with this model which is better than Random forest model but not as good as Logistic Regression.

Frequent Q&A

Q) What is the source of the data?

- Data was collected from VIEH group GItHub repository.

Q) What is the complete flow of your project?

- Refer to slide no 4 for better understanding.

Q) What techniques were you using for data pre-processing?

- In data pre processing, we analyzed the data, found the important features, and based on the domain knowledge, we eliminated the unnecessary columns. We also tried to deal with missing value (in our case there are no missing value). We also deal with the categorical data.

Frequent Q&A

Q) How did you choose the model?

- After implementing 3 models (logistic regression, random forest, support vector machine), we were able to do model selection based on the high accuracy of a particular model. The final model we chose was Logistic Regression.