Machine Learning Project -

Predicting Employee Attrition

*Business Objective*

An IT company has a workforce of around 4,000 employees. Every year around 15% of the employees leave the company and need to be replaced with new hires. The leadership believes that 15% attrition is bad for the company’s revenue and future growth. The Company has approached you to build a machine learning model that could predict how likely it is for an employee to leave the company. The model should guide the company to take timely action to retain as many employees as possible.

*Problem Statement*

You have been assigned the task to build for this company a Decision Tree Machine Learning model that can predict the probable attrition of the employees. *For more details on the intermediate and final outputs expected, refer to the list of deliverables mentioned in the “Model Building” and “Model Validation” sections below.*

*Data Description*

The dataset provided for this activity consists of 30 features where 29 are independent features and 1 is a target variable. Features in this dataset are described as below :

| **Index** | **Variable** | **Meaning** |
| --- | --- | --- |
| 1 | Age | Age of the employee |
| 2 | Attrition | Whether the employee left in the previous year or not |
| 3 | BusinessTravel | How frequently the employees travelled for business purposes in the last year |
| 4 | Department | Department in company |
| 5 | DistanceFromHome | Distance from home in km |
| 6 | Education | Education Level |
|
|
|
|
|
| 7 | EducationField | Field of education |
| 8 | EmployeeCount | Employee count |
| 9 | Employee number | Employee number/id |
| 10 | EnvironmentSatisfaction | Work Environment Satisfaction Level |
|
|
|
| 11 | Gender | Gender of employee |
| 12 | JobInvolvement | Job Involvement Level |
|
|
|
| 13 | JobLevel | Job level at company on a scale of 1 to 5 |
| 14 | JobRole | Name of job role in company |
| 15 | JobSatisfaction | Job Satisfaction Level |
|
|
|
| 16 | MaritalStatus | Marital status of the employee |
| 17 | MonthlyIncome | Monthly income in rupees per month |
| 18 | NumCompaniesWorked | Total number of companies the employee has worked for |
| 19 | Over18 | Whether the employee is above 18 years of age or not |
| 20 | PercentSalaryHike | Percent salary hike for last year |
| 21 | PerformanceRating | Performance rating for last year |
|
|
|
| 22 | RelationshipSatisfaction | Relationship satisfaction level |
|
|
|
| 23 | StandardHours | Standard hours of work for the employee |
| 24 | StockOptionLevel | Stock option level of the employee |
| 25 | TotalWorkingYears | Total number of years the employee has worked so far |
| 26 | TrainingTimesLastYear | Number of times training was conducted for this employee last year |
| 27 | WorkLifeBalance | Work life balance level |
|
|
|
|
| 28 | YearsAtCompany | Total number of years spent at the company by the employee |
| 29 | YearsSinceLastPromotion | Number of years since last promotion |
| 30 | YearsWithCurrManager | Number of years under current manager |

*Model Building*

* Show Bi-variate plots (scatter/ bar) of all meaningful variables with the dependent variable
* For the root node, show the Gini value for all the categorical variables?
* Show (using relevant metrics) how much overfitting did you observe with the Raw (unpruned) model?
* Summarize the steps followed to finalize your model - consisting of the below steps (as applicable)
  + Sampling
  + Feature Engineering
  + Performance comparison between Train and Test
  + Use of Cross-validation
* While developing the model, you would have gotten a few candidate models which were not as good as the final model (in terms of performance, multicollinearity, or statistical stability etc.). Show a few of these candidate models and explain their shortcomings
* What approach did you follow in pruning your decision tree model? Which Hyperparameter(s) did you choose and why? Show relevant metrics/output to substantiate your approach.
* Show what kind of feature engineering did you apply in your project and why (include in your results what’s applicable from below)
  + Dummy variables
  + Label encoding
  + Any bin-based variable created -what was the significance/rationale of binning
  + Any new derived variables created using the raw variables – For e.g., Ratio based, difference based, % difference based / Rate of change, etc.
* If the provided dataset is unbalanced, what steps did you take to balance it. Also, explain the technique used to oversample/under-sample the dataset?
* For a few important independent variables, show how the decision boundary looks like for the finalized model? Hint: At a time, you will need to choose a combination of two independent variables.
* Demonstrate Live how your model will assign class/ or compute the probability for a new data point?
* Provide your understanding of the next steps that the client/ end-user needs to follow to deploy your model at their end. Think about the below lines:
  + Any technical/infrastructure requirements that the client needs to meet?
  + What files do you need to provide them?
  + What kind of data cleaning and preprocessing would the client need to do before using the model?
  + How will the client use your model on new data?
  + How will the client know that the model is performing well on new data points?

Model Validation

* Show your model’s performance on the below metrics (on both train and test samples)
  + Confusion Matrix
  + Classification Report
  + Concordance test results (This coefficient is used to assess the agreement between estimated values and correct values)
  + Rank ordering test results (Rank ordering is an important measure of model performance and its ability to separate out the event from the non-events)
* For the given business problem which of the below metric(s) did you choose and why? Include in your final output any additional activity performed (and its results) to get to the best values of the below metrics (F1-Score, AUC-ROC curve, AUC-ROC Accuracy).
  + Accuracy
  + Precision
  + Recall