Employee Turnover Prediction

Overview of this document:

As part of the “Employee Turnover Prediction” project this documents contains information about what are the data sources are used, how they are helped , used dataset link and also include what the required libraries to implement the project.

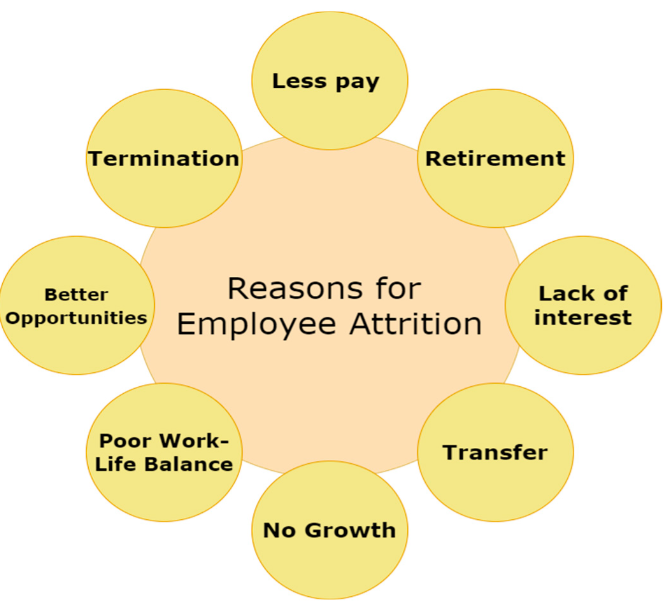
Data Sources :

Based on the below sources the project will be going to implement in effective manner

* Data Link: <https://www.mdpi.com/2673-4591/59/1/117>

The above link is from “**MDPI”** website. This contains Proceeding paper of Predicting Employee Turnover. The terms used in this is very easy to understand and implement the process. That proceeding paper presents a discussion of the numerous approaches that may be taken to forecast employee turnover, and it also includes an analysis of the most effective solution that was conducted by comparing different models.

The various reasons why an employee may decide to leave an organization:



Reasons for employee attrition

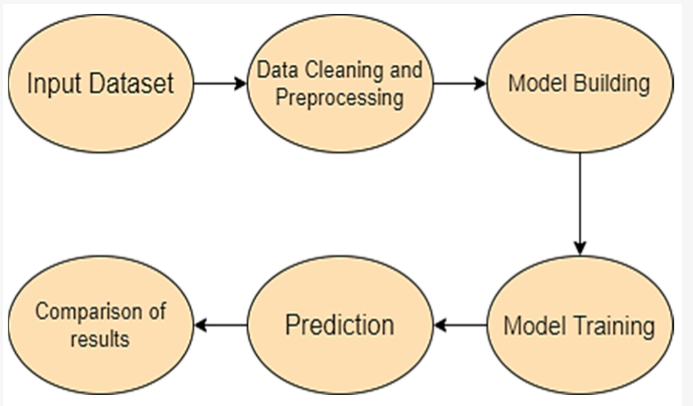
* Data Link: https://www.researchgate.net/publication/328772915\_Employee\_Turnover\_Prediction\_with\_Machine\_Learning\_A\_Reliable\_Approach

The above link is taken from “**ResearchGate”** website, published by University of Toronto, Toronto, Canada. It provides a detailed explanation about what are the technologies to be used and how to implement.

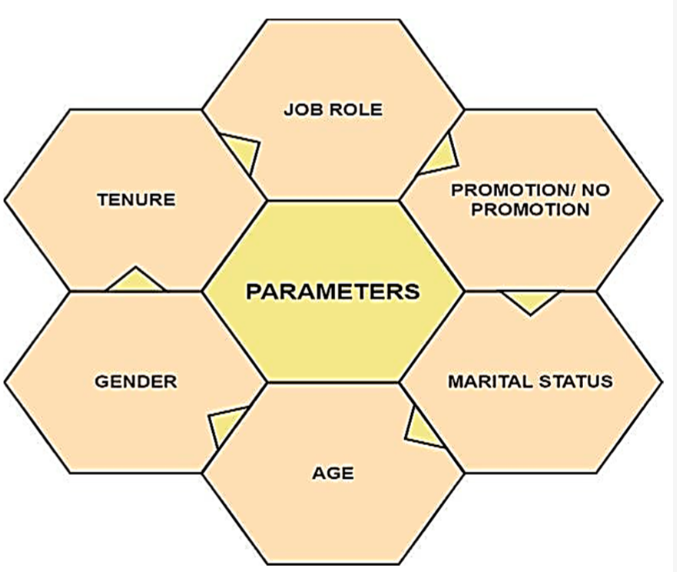
 Methodology:

*Design, Architecture, and Dataset*

As per the architecture shown in the below figure, the system that has been proposed was built using several different types of machine learning. Each model makes use of the same dataset to make a prediction regarding attrition. The collection is made up of a variety of different personnel records (both past and present) . The incoming dataset is initially subjected to cleaning and preprocessing, which involves the management of all missing values, NaN values, etc., as well as the removal of unwanted columns. After that follows the process of model construction, which involves considering several different models to make a prediction. The dataset is then divided into a training dataset and a test dataset, with the training dataset being the one that is utilized to train each model employed. Following a comparison of all the predictions based on the evaluation measures, the most effective model is proposed.



Machine learning architecture



Features used in prediction

Libraries required:

Employee turnover prediction typically involves using various libraries and tools depending on the approach and technology stack chosen by the data scientists or analysts. Here are some common libraries and tools used in the process:

**Python Libraries**:

* **Pandas**: For data manipulation and preprocessing.
* **NumPy**: For numerical computations.
* **scikit-learn**: For machine learning algorithms such as logistic regression, decision trees, random forests, etc.
* **TensorFlow or PyTorch**: For building and training deep learning models, if applicable.
* **Statsmodels**: For statistical models and tests.
* **Seaborn and Matplotlib**: For data visualization.

**R Libraries** (less common but still used in some cases):

* **caret**: For machine learning workflows.
* **glmnet**: For regularized regression models.
* **randomForest**: For random forest models.
* **ggplot2**: For data visualization

**Specific Packages for Turnover Prediction**:

* **imbalanced-learn**: For dealing with imbalanced datasets (if turnover is relatively rare).
* **XGBoost or LightGBM**: For gradient boosting models, which can sometimes outperform traditional models.
* **Feature-engine**: For feature engineering tasks like handling missing data, encoding categorical variables, etc.

**Feature Selection and Engineering**:

* **Featuretools**: Automated feature engineering based on relational data.
* **Scikit-plot**: Easy visualization of model performance and feature importance.

**Deployment and Monitoring**:

* **Flask or FastAPI**: Lightweight frameworks for deploying machine learning models as APIs.
* **Docker and Kubernetes**: For containerization and orchestration of model deployment.
* **Prometheus and Grafana**: For monitoring deployed models and infrastructure.

**Additional Tools**:

* **SQL and database management tools**: For data extraction and storage.
* **Excel or Google Sheets**: For initial data exploration and simple analysis.
* **Jupyter Notebooks or RStudio**: For interactive development and analysis.
* **Git**: For version control of code and models.

Conclusion:

In conclusion, an employee turnover prediction project typically involves leveraging a variety of libraries and tools to analyze historical data and forecast future turnover. Key libraries include foundational tools like Pandas, NumPy, scikit-learn, and TensorFlow/PyTorch for data manipulation, modeling, and potentially deep learning. Specialized tools such as imbalanced-learn, XGBoost, and SHAP enhance model performance and interpretability. Additional considerations like NLP for text analysis, time series techniques for temporal data, and AutoML for automated model selection may also be employed based on project requirements. Successful execution of such a project requires careful data preprocessing, model selection, and interpretation of results to provide actionable insights for mitigating employee turnover and optimizing workforce management strategies.