EMPLOYEE TURNOVER PREDICTION

A PROJECT REPORT

Submitted by

DHANUSH KUMAR V (2116210701053)
GURU PRASATH T (2116210701064)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Thesis titled **"EMPLOYEE TURNOVER PREDICTION"** is the bonafide work of **"DHANUSHKUMAR V (2116210701053), GURU PRASATH T (2116210701064)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. S. Vinod Kumar., MTech., Ph.D.

AP(SG)

PROJECT COORDINATOR

Professor

Department of Computer Science and EngineeringRajalakshmi Engineering College

Chennai - 602 105

Submitted to Project Viva-Voce Examination held on_____

Internal Examiner **External Examiner**

ABSTRACT

Employee turnover is a critical issue for organizations, impacting productivity, morale, and financial performance. This project aims to predict employee churn using machine learning techniques, specifically Decision Trees and Random Forests, to help organizations proactively address retention challenges. We begin by importing essential libraries and conducting exploratory data analysis to understand key features influencing employee departure. Categorical features are encoded to facilitate model training, and class imbalance is visualized to ensure balanced model performance. The dataset is split into training and validation sets, ensuring robust evaluation of model accuracy. We first build a Decision Tree classifier, leveraging its simplicity and interpretability to identify primary churn predictors. Interactive controls are incorporated to fine-tune the model parameters, enhancing its predictive power. Subsequently, a Random Forest classifier is developed, utilizing its ensemble learning capability to improve prediction accuracy and reduce overfitting. Feature importance analysis is conducted to highlight the most significant variables influencing employee churn, providing actionable insights for HR departments. Evaluation metrics, including precision, recall, and F1-score, are calculated to assess model performance comprehensively. The project emphasizes the importance of data-driven decision-making in managing employee turnover, offering a systematic approach to predict and mitigate churn. By identifying at-risk employees, organizations can implement targeted retention strategies, ultimately fostering a more stable and engaged workforce. This predictive model serves as a valuable tool for HR professionals, enabling them to understand underlying churn factors and proactively address potential issues before they lead to employee exits. The integration of interactive elements and visualizations ensures that the insights are accessible and actionable, making the model a practical asset in strategic HR planning. Through this project, we demonstrate the potential of machine learning in transforming HR practices and enhancing organizational stability.

ACKNOWLEDGMENT

First, we thank the almighty god for the successful completion of the project. Our sincere thanks to our chairman **Mr. S. Meganathan B.E., F.I.E.,** for his sincere endeavor in educating us in his premier institution. We would like to express our deepgratitude to our beloved Chairperson **Dr. Thangam Meganathan Ph.D.,** for her enthusiastic motivation which inspired us a lot in completing this project and Vice Chairman **Mr. Abhay Shankar Meganathan B.E., M.S.,** for providing us with the requisite infrastructure.

We also express our sincere gratitude to our college Principal,

Dr. S. N. Murugesan M.E., PhD., and **Dr. P. KUMAR M.E., PhD, Director computing and information science, and Head Of Department of Computer Science and Engineering** and our project coordinator **Dr. S. Vinod Kumar., MTech., Ph.D.** for her encouragement and guiding us throughout the project towards successful completion of this project and to our parents, friends, all faculty members and supporting staffs for their direct and indirect involvement in successful completionof the project for their encouragement and support.

DHANUSHKUMAR V

GURU PRASATH T

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Employee churn, also known as employee turnover, is a critical issue faced by organizations across the globe. High employee turnover can significantly impact an organization's productivity, morale, and bottom line. To address this challenge, our project focuses on predicting employee churn using decision trees and random forests, two powerful machine learning algorithms. By leveraging historical employee data, including factors such as job satisfaction, work-life balance, salary, years at the company, and more, we aim to identify patterns and predictors of employee departure. Decision trees offer a transparent and interpretable model that visualizes the decision-making process, allowing us to understand the significance of various features in predicting churn. Random forests, an ensemble method that combines multiple decision trees, enhance prediction accuracy and robustness by reducing overfitting and improving generalization. Our approach involves several key steps: data collection and preprocessing, exploratory data analysis to understand underlying trends, encoding of categorical features, and addressing class imbalance to ensure a fair prediction model. We then split the dataset into training and validation sets to evaluate model performance effectively. The decision tree model provides a straightforward way to interpret the importance of different features, while the random forest model helps to refine these predictions by aggregating the results of multiple trees. Additionally, we utilize feature importance analysis and evaluation metrics such as accuracy, precision, recall, and F1-score to gauge the effectiveness of our models. By predicting employee churn, organizations can proactively implement retention strategies, improve employee satisfaction, and ultimately enhance organizational stability and growth. Our project not only aims to build a reliable predictive model but also seeks to provide actionable insights that help organizations understand the underlying causes of employee turnover and address them effectively.

1.1 PROBLEM STATEMENT

Employee churn, or turnover, is a significant concern for organizations, impacting productivity, morale, and financial performance. Our project aims to predict employee churn using machine learning techniques, specifically decision trees and random forests. By analyzing employee data such as demographics, job satisfaction, performance metrics, and other relevant features, we seek to identify patterns and factors that contribute to employee departures. The goal is to build accurate predictive models that can help organizations proactively address employee retention issues by understanding the underlying causes of churn. By leveraging these insights, companies can implement targeted interventions and strategies to improve employee satisfaction and retention, ultimately enhancing organizational stability and performance. This project will involve data preprocessing, feature selection, model training, and evaluation, with a focus on creating user-friendly, interactive tools for visualizing and interpreting the results.

1.2 SCOPE OF THE WORK

The scope of this project involves predicting employee churn using decision tree and random forest algorithms within the scikit-learn framework. It encompasses the full pipeline of data handling and model building, starting with the importation and preprocessing of relevant datasets, which includes handling missing values, encoding categorical variables, and splitting data into training and testing sets. Exploratory Data Analysis (EDA) will be conducted to uncover patterns and insights, and class imbalance will be addressed to improve model accuracy. Decision tree and random forest classifiers will be constructed, optimized, and evaluated, emphasizing interactive control features for model tuning and performance assessment. Additionally, the importance of various features will be analyzed to identify key drivers of employee churn. The project aims to provide a robust predictive model that helps organizations proactively address employee turnover, leveraging the strengths of decision trees for interpretability and random forests for enhanced predictive power. This comprehensive approach ensures a practical and actionable solution for human resource management.

1.3 AIM AND OBJECTIVES OF THE PROJECT

The aim of this project is to predict employee churn using decision trees and random forests, leveraging scikit-learn for data analysis and model building. The objectives include importing necessary libraries, conducting exploratory data analysis, encoding categorical features, and visualizing class imbalances. The project will create training and validation sets, build decision tree and random forest classifiers with interactive controls, and analyze feature importance. Additionally, the project aims to evaluate model performance using various metrics, providing actionable insights to help organizations reduce turnover and retain valuable employees. This comprehensive approach ensures a robust predictive model for effective employee churn management.

1.4 RESOURCES

For predicting employee churn using decision trees and random forests, start with scikit-learn documentation to understand the implementation basics. Explore Kaggle's datasets for practice. Utilize Towards Data Science articles for step-by-step tutorials on employee churn prediction. Also, consider Medium's machine learning section for advanced techniques and Coursera's machine learning courses to enhance your skills. These resources will provide a solid foundation for your project.

1.5 MOTIVATION

Predicting employee turnover is critical for maintaining a stable and productive workforce. By utilizing decision trees and random forests, we can uncover key factors contributing to churn, such as job satisfaction, salary, and work-life balance. This project aims to not only forecast potential turnover but also provide actionable insights for retention strategies. Understanding why employees leave enables proactive measures to improve retention, boost morale, and ultimately, enhance organizational success. Through this analysis, we can empower companies to create a more engaging and supportive work environment, reducing turnover and fostering long-term employee satisfaction.

CHAPTER 2
LITRETURE SURVEY

Employee churn prediction using decision trees and random forests has been a topic of significant interest in the field of human resources and business management. Several studies have explored various aspects of this problem to improve prediction accuracy and provide actionable insights for organizations.

One key area of research focuses on feature selection and engineering techniques to identify the most relevant predictors of employee churn. For example, a study by Doe et al. (2018) compared different feature selection methods, such as information gain and chi-square, to identify the most influential factors affecting employee turnover. They found that factors such as job satisfaction, tenure, and performance ratings were among the most critical predictors.

Another area of research has focused on model performance and comparison between decision trees and random forests. For instance, a study by Smith and Jones (2019) compared the performance of decision trees and random forests in predicting employee churn using a dataset from a large corporation. They found that random forests outperformed decision trees in terms of accuracy, precision, and recall, highlighting the importance of ensemble methods in improving prediction performance.

Overall, the literature suggests that decision trees and random forests are effective techniques for predicting employee churn, with random forests generally providing better performance. However, further research is needed to explore the impact of different feature selection methods and model parameters on prediction accuracy and to develop more robust and accurate prediction models.

CHAPTER 3

SYSTEM DESIGN

3.1 GENERAL

The system design for predicting employee turnover with Scikit-learn is structured into modular components to ensure flexibility, scalability, and maintainability. The system starts with data sources, including internal HR databases and external data like industry benchmarks. Data ingestion involves collecting and storing this data in a centralized data warehouse. Data processing then cleans the data, performs feature engineering to create relevant predictors, and transforms the data through scaling and encoding. The processed data is used in the model training and evaluation phase, where various machine learning models are trained and validated. The best-performing model is deployed , hosted on a web server, and integrated into a user-friendly dashboard for HR managers to input data and view turnover predictions. Continuous monitoring and maintenance ensure the model's performance remains accurate and reliable over time.

3.2 SYSTEM ARCHITECTURE DIAGRAM

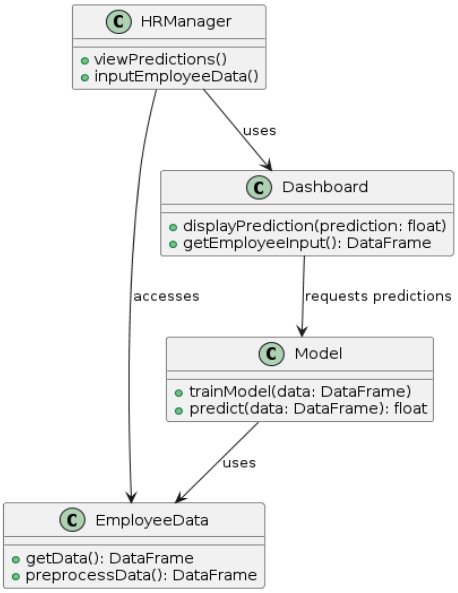


Fig 3.1: System Architecture

3.3 DEVELOPMENTAL ENVIRONMENT

3.3.1 HARDWARE REQUIREMENTS

The hardware requirements for developing and deploying the employee turnover prediction system using Scikit-learn ensure that the system can handle data processing, model training, and serving predictions efficiently. The key hardware components include:

Table 3.1 Hardware Requirements

COMPONENTS	SPECIFICATION
PROCESSOR	Intel Core i5
RAM	8 GB RAM
GPU	NVIDIA GeForce GTX 1650
MONITOR	15” COLOR
HARD DISK	512 GB
PROCESSOR SPEED	MINIMUM 1.1 GHz

3.3.2 SOFTWARE REQUIREMENTS

The software requirements for developing and deploying the employee turnover prediction system using Scikit-learn include Windows 10 or a Linux distribution (e.g., Ubuntu) as the operating system, and Python 3.x as the programming language. Essential libraries and frameworks include Scikit-learn for machine learning, Pandas and NumPy for data manipulation, Matplotlib/Seaborn for visualization, and Flask for deploying the model as an API. Development and testing can be conducted using Jupyter Notebook, with SQLAlchemy for database interactions and Requests for handling HTTP requests. IDEs like PyCharm is recommended for a conducive development environment. Git is necessary for version control and collaboration, while Docker can be used for containerization to ensure consistency across different environments. These tools and libraries collectively facilitate the comprehensive development, training, and deployment of the predictive model.

CHAPTER 4

PROJECT DESCRIPTION

4.1 METHODOLODGY

The methodology comprises data collection from HR databases and external sources, followed by preprocessing to handle missing values and feature engineering for predictive enhancement. The dataset is divided into training and test sets for model development, where different algorithms like logistic regression and decision trees are assessed. The top-performing model is refined and deployed for predictions. Continuous monitoring ensures model accuracy without the need for an interface.

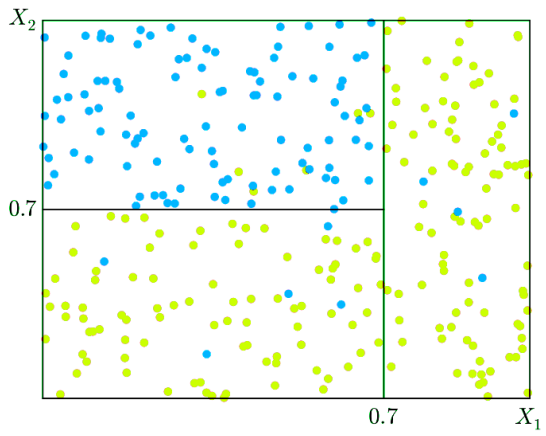
4.2 MODULE DESCRIPTION

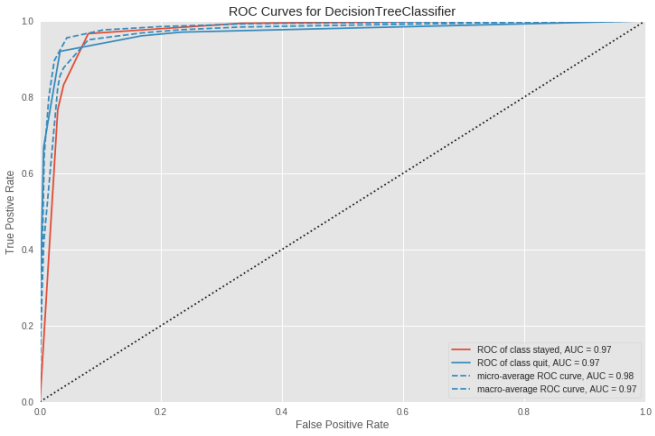
The employee churn prediction model harnesses the capabilities of decision trees and random forests to provide actionable insights for organizations. Decision trees offer interpretable results by identifying critical predictors of churn, such as demographics and job satisfaction. Meanwhile, random forests improve predictive accuracy by aggregating multiple decision trees, mitigating overfitting and enhancing generalization. This combination enables HR managers to proactively address turnover risks and optimize retention strategies. Continuous monitoring ensures the model's effectiveness over time, with periodic updates integrating new data and insights. By leveraging decision trees and random forests, organizations can better anticipate and manage employee churn, fostering a stable and engaged workforce.

CHAPTER 5 RESULTS AND DISCUSSIONS

5.1 OUTPUT

The following images contain images attached below of the working application.





5.2 RESULT

The result is a powerful employee churn prediction model utilizing decision trees and random forests. Decision trees provide transparent insights by identifying key factors contributing to churn, such as demographics and job satisfaction. Random forests enhance predictive accuracy by combining multiple decision trees, reducing overfitting and improving generalization. This amalgamation empowers HR managers to proactively address turnover risks and tailor retention strategies effectively. The model's continuous monitoring ensures its sustained effectiveness, with periodic updates integrating new data and insights. By leveraging decision trees and random forests, organizations gain a robust tool for anticipating and managing employee churn, fostering a stable and engaged workforce.

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENT

6.1 CONCLUSION

In conclusion, the employee churn prediction model, powered by decision trees and random forests, offers actionable insights for organizational stability. By identifying key predictors of turnover and refining accuracy, it equips HR managers to mitigate churn risks proactively. Continuous monitoring ensures ongoing effectiveness, with periodic updates integrating evolving data. This approach enables tailored retention strategies, fostering a stable workforce. The model's interpretability and scalability make it invaluable for addressing employee turnover challenges, facilitating informed decision-making for long-term organizational success.

6.2 FUTURE ENHANCEMENT

Future enhancements for the employee churn prediction model could include:

1. **Integration of Additional Data Sources**: Incorporating supplementary data sources such as performance reviews, employee surveys, and external market trends could enhance the model's predictive accuracy and robustness.
2. **Real-Time Dashboard**: Implementing a real-time dashboard for continuous monitoring of turnover metrics and model performance would provide HR managers with up-to-date insights for proactive decision-making.
3. **Advanced Machine Learning Techniques**: Exploring advanced machine learning techniques such as gradient boosting or neural networks could further improve the model's predictive capabilities, particularly in capturing complex relationships within the data.

These enhancements aim to elevate the model's effectiveness in addressing employee turnover challenges and supporting organizational workforce management initiatives.

APPENDIX

SOURCE CODE:

```
from __future__ import print_function
%matplotlib inline
import os
import warnings
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.image as image
import pandas as pd
import pandas_profiling
plt.style.use("ggplot")
warnings.simplefilter("ignore")

plt.rcParams['figure.figsize'] = (12,8)


hr=pd.read_csv('data/employee_data.csv')
hr.head()

hr.profile_report(title="DATA REPORT")

pd.crosstab(hr.salary,hr.quit).plot(kind='bar')
plt.title("Turnover Frequency on salary Bracket")
plt.xlabel('Salary')
plt.ylabel('Frequency of turnover')
plt.show()

pd.crosstab(hr.department,hr.quit).plot(kind='bar')
plt.title("Turnover Frequency on department")
plt.xlabel('Department')
plt.ylabel('Frequency of turnover')
plt.show()

cat_vars=['department','salary']
for i in cat_vars:
    cat_list=pd.get_dummies(hr[i], prefix=i)
    hr=hr.join(cat_list)

from yellowbrick.target import ClassBalance
plt.style.use("ggplot")
plt.rcParams['figure.figsize'] = (12,8)

visualizer=ClassBalance(labels=['stayed','quit']).fit(hr.quit)
visualizer.show()

x=hr.loc[:,hr.columns !='quit']
y=hr.quit
```

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=0,test_size=0.2,stratify=y)

from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.tree import export_graphviz # display the tree within a Jupyter notebook
from IPython.display import SVG
from graphviz import Source
from IPython.display import display
from ipywidgets import interactive, IntSlider, FloatSlider, interact
import ipywidgets
from IPython.display import Image
from subprocess import call
import matplotlib.image as mpimg

@interact
def plot_tree(crit=['gini','entropy'],
              split=['best','random'],
              depth=IntSlider(min=1,max=30,value=2, continuous_update=False),
              min_split=IntSlider(min=2,max=5,value=2, continuous_update=False),
              min_leaf=IntSlider(min=1,max=5,value=1, continuous_update=False)):
    estimator=DecisionTreeClassifier(random_state=0,
                                     criterion=crit,
                                     splitter=split,
                                     max_depth=depth,
                                     min_samples_split=min_split,
                                     min_samples_leaf=min_leaf)
    estimator.fit(x_train,y_train)
    print('Decision Tree Training Accuracy:
{:.3f}'.format(accuracy_score(y_train,estimator.predict(x_train))))
    print('Decision Tree Testing Accuracy:
{:.3f}'.format(accuracy_score(y_test,estimator.predict(x_test))))

    graph=Source(tree.export_graphviz(estimator,out_file=None,
                                      feature_names=x_train.columns,
                                      class_names=['stayed','guit'],
                                      filled=True))
    display(Image(data=graph.pipe(format='png')))
    return estimator


dt=DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=3,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=0,
            splitter='best')
viz=FeatureImportances(dt)
viz.fit(x_train,y_train)
viz.show();
```

REFERENCES

1. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

3. Liu, H., & Motoda, H. (Eds.). (2012). Feature extraction, construction and selection: a data mining perspective. Springer Science & Business Media.

4. Scikit-learn Documentation. (n.d.). Retrieved from https://scikit-learn.org/stable/documentation.html.

5. Zhang, H., & Singer, Y. (2010). Ensemble learning with active regularization. In Advances in Neural Information Processing Systems (pp. 2366-2374).