Employee Attrition & Financial Impact Analysis Report

Overview

This project aims to help organizations better understand and act upon employee attrition by combining classification and regression tasks. The objective is to not only predict who might leave but also simulate and predict the financial consequences of attrition. By doing so, we offer companies data-driven insights to prioritize retention efforts.

Project Goals:

1. Predict employee attrition (classification)
2. Simulate and predict future salary of likely-to-stay employees (regression)
3. Estimate potential financial loss due to attrition

Dataset Used:

IBM HR Analytics Attrition Dataset from Kaggle

Data Preprocessing

Before feeding data into the models, several preprocessing steps were necessary:

- **Handling Categorical Data:**
  Categorical features such as 'Department', 'Gender', and 'JobRole' were encoded using LabelEncoder, which assigns a unique integer value to each category. This transformation allows the models to interpret these non-numeric features.
- **Scaling Numeric Features:**
  Features like 'Age', 'DistanceFromHome', and 'MonthlyIncome' were scaled using StandardScaler to standardize the data. This is especially important for algorithms like SVM and Logistic Regression, which are sensitive to the scale of input features.
- **Simulating Future Salary:**
  Since the dataset lacked actual future salary data, we simulated it based on performance and salary:
  - High performers (rating = 4) received a 10% increment.
  - Others received a 5% increment.

This was necessary for training our regression model, as future salary prediction is an essential component of calculating the financial impact of employee attrition.

Classification Models (Attrition Prediction)

To predict employee attrition, several classification models were trained:

- **Logistic Regression:**
  As a baseline model for binary classification, Logistic Regression is simple, interpretable, and suitable for linear relationships.
- **Decision Tree Classifier:**
  This model was chosen to capture non-linear relationships in the data, particularly where decision boundaries between classes are not straight lines.
- **Support Vector Classifier (SVC):**
  SVC was used for its ability to handle complex, high-dimensional data by mapping it to higher dimensions where classes can be separated by a hyperplane.

Best Performing Model: Logistic Regression

- Logistic Regression was chosen as the best model due to its balance between performance and interpretability. Although it may not perform as well as more complex models in some cases, its coefficients provide clear insight into which features are most important for predicting attrition.

Evaluation Metrics:

- F1 Score
- AUC-ROC
- Precision, Recall
- 5-Fold Cross Validation

Classification Report

- F1 Score: 0.503
- AUC-ROC: 0.796
- Precision: 0.375
- Recall: 0.766
- Cross-Validation F1 Score: 0.489

The model performed reasonably well, indicating that while we have some false positives, our recall (identifying employees likely to leave) is relatively strong.

Identifying Likely-to-Stay Employees

In order to predict future salary, we needed to identify employees likely to stay at the company:

- **Calculated Probability of Staying:**
  The probability of staying is simply the complement of the probability of attrition:
  P(Stay) = 1 - P(Attrition)
- **Selection Criteria:**
  Employees with a P(Stay) > 0.6 were selected. This threshold was chosen to focus on employees with a high likelihood of staying and to ensure the regression model is only trained on employees whose future salaries we are more confident about.

Regression Models (Future Salary Prediction)

To predict the future salary for employees who are likely to stay, the following regression models were considered:

- **Random Forest Regressor:**
  Random Forest is an ensemble method that builds multiple decision trees and combines their predictions. It is particularly effective for capturing complex relationships and feature interactions without requiring explicit modeling of the underlying data distribution.
- **Ridge Regression:**
  Ridge Regression adds a penalty to the size of coefficients to prevent overfitting. It was considered because it can handle multicollinearity by shrinking large coefficients, though it performs better with less complex datasets.
- **Lasso Regression:**
  Lasso Regression performs feature selection by forcing some coefficients to zero. It was useful for identifying the most important features for predicting future salary.
- **Support Vector Regressor (SVR):**
  SVR was tested to capture non-linear relationships between features and salary. However, it can be sensitive to the scaling of input features and is less effective when the relationships are very complex.

Best Performing Model: Random Forest Regressor

- R^2 Score: 0.9999
- RMSE: 30.54

The Random Forest model outperformed others due to its ability to automatically handle feature interactions and non-linear relationships in the data. This makes it a robust choice for salary prediction, providing both high predictive accuracy and a strong ability to generalize.

Other Models:

- **Ridge Regression:**
  - R^2 Score: 0.9159
  - RMSE: 371.13
- **Lasso Regression:**
  - R^2 Score: 0.9145
  - RMSE: 373.31
- **SVR:**
  - R^2 Score: 0.2428
  - RMSE: 1084.25

The SVR model underperformed due to its sensitivity to scaling and failure to capture the complex, non-linear relationships in the salary data. Ridge and Lasso performed decently but still didn't match the Random Forest model's performance.

Financial Impact Estimation

The key part of this analysis is estimating the financial impact of attrition. We calculated the expected financial loss for each employee due to attrition using the following formula:

- **Expected Loss per Employee = Probability of Attrition × Predicted Future Salary**
  Example:
  Expected Loss (Employee i) = P(Attrition for Employee i) × FutureSalary(i)

To aggregate the total impact across all employees, we summed the individual expected losses:

- **Total Expected Financial Loss = Sum of (Expected Loss for each employee)**

This provides a quantifiable estimate of the financial impact of employee turnover.

Financial Loss Estimate

- **Total Expected Financial Loss: $ 2,934,964.70**

This amount represents the total value that the company stands to lose due to employees predicted to leave, based on their future salary and probability of attrition.

Conclusion

- This multi-step project successfully modeled real-world HR decision-making by:
- Identifying potential attrition risks.
- Estimating their financial value to the company.
- Highlighting who should be prioritized for retention strategies.

Why Random Forest Was Best:

- Non-linear Relationships: Random Forest is excellent at capturing non-linear relationships in salary data.
- No Assumptions: It does not require assumptions like linearity or normality, making it versatile across different datasets.
- Strong Predictive Performance: The Random Forest model showed near-perfect R^2 and very low RMSE, making it a reliable predictor.

Future Improvements:

- More Features: Include additional data such as peer review scores, bonus history, or project load to improve model performance.
- Model Interpretability: Use techniques like SHAP or feature importance to explain which features drive attrition and salary growth.
- Ensemble Methods: Implementing ensemble stacking could combine the strengths of different models for even better performance.

This project bridges data science and HR strategy—supporting both data-driven decisions and human insight.

Prepared using Python, Scikit-learn, Pandas, and Matplotlib.