**1. Analysis**

**1.1 Research Questions**

● What kind of people tend to have higher salary? What factors have the greatest impact on the monthly income? Can we predict the monthly income based on other features of the employees? We will fit, interpret, and evaluate the random forest model for monthly income.

● Which factors have the greatest impact on the attrition? Can we predict the attrition of employees based on other characteristics? We will use logistic regression model and Support Vector Machines (SVM) to build prediction models for attrition and select the best method with the highest accuracy.

**1.2 Exploratory Data Analysis**

**1.2.1 Dataset Description**

The “IBM HR Analytics Employee Attrition & Performance” dataset is from the Kaggle website(https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset?select=WA\_Fn-UseC\_-HR-Employee-Attrition.csv), it is used to train a machine learning model to predict attrition of the valuable employees based on a list of attributes.

The structure of the dataset is shown in Figure 1. The dataset has a total of 1470 rows and 35 columns, of which 9 are numerical variables and 26 are categorical variables. Since each row represents several attributes of an employee from IBM, the unit of analysis in this dataset would be a given employee.

表格

描述已自动生成

***Figure. 1. Structure of the dataset***

Some attributes need further explanations for better understandings are shown in Table 1.

***Table. 1. Explanations of the variables***

|  |  |
| --- | --- |
| **Variable** | **Explanation** |
| Education | 1 'Below College', 2 'College', 3 'Bachelor', 4 'Master', 5 'Doctor' |
| EnvironmentSatisfaction | 1 'Low', 2 'Medium', 3 'High', 4 'Very High' |
| JobInvolvement | 1 'Low', 2 'Medium', 3 'High', 4 'Very High' |
| JobSatisfaction | 1 'Low', 2 'Medium', 3 'High', 4 'Very High' |
| PerformanceRating | 1 'Low', 2 'Good', 3 'Excellent', 4 'Outstanding' |
| RelationshipSatisfaction | 1 'Low', 2 'Medium', 3 'High', 4 'Very High' |
| WorkLifeBalance | 1 'Bad', 2 'Good', 3 'Better', 4 'Best' |

**1.2.2 Data preparation**

First, we found that there are no empty or duplicated records or missing data in the dataset. There are some outliers in variables ‘MonthlyIncome’, ‘NumCompaniesWorked’, etc. Outliers make the data more variable, which reduces statistical power. As a result, eliminating outliers might make the findings statistically significant. So, we removed these outliers for better predictions. Also, we removed three variables ‘Over18’, ‘EmployeeCount’, and ‘EmployeeNumber’ as they are useless for the prediction. All employees are over 18 years old, and all employees have an "EmployeeCount" value of 1. 'EmployeeNumber' is only used to identify the employee and has no meaning for model predictions.

Finally, we encode the categorical variables for model predictions. For variables like ‘Gender’, we use 1 and 0 to represent ‘female’ and ‘male, while for variables like ‘EducationField’ which has multiple factors, we used one-hot encoding method to encode them.

**1.2.3 Analysis and Interpretation**

After the cleaning, we got 46 numeric variables and 1,435 records in the dataset.

In the correlation plot shown in Figure 2, “JobLevel’ has the highest correlation of 0.95 with the monthly income, followed by “TotalWorkingYears” of 0.77.

Then we checked the relationship between ‘MonthlyIncome’ and these two varables. From Figure 3, we can see that employees with higher job level tend to have higher salary. From Figure 4, we can see that the longer the working life, the higher the salary.

图表, 散点图

描述已自动生成

***Figure. 2. Correlation Matrix plot for numeric variables***

图表, 箱线图

描述已自动生成

***Figure. 3. Distribution of monthly income by job level***

图表, 散点图

描述已自动生成

***Figure. 4. Monthly income vs Total working years***

**1.3 Random Forest for Predicting Monthly Income**

**1.3.1 Method and Reasons:**

We performed Random Forest method to predict the Monthly Income of the employees. The reasons are shown as below:

Random forest regression yields incredibly high accuracy thanks to its "wisdom of the crowds" strategy. Compared to other linear models, such as logistic regression and linear regression, it typically yields better results. The approach scales effectively in terms of computation when new characteristics or samples are introduced to the dataset. It is not overly sensitive to missing data, and it has some tolerance for outliers.

**1.3.2 Analysis Steps:**

To assess the model's predictive power, we must randomly divide the data into a training group and a test group and set the data scale ratio of the two to 7:3.

We used the randomForest function to build a random forest model to predict price. As shown in Figure 5, % Var explained represents the overall explained rate of variance related to the response variable for the predictor variables. In this model, about 94.46% of the total variance were explained, which can be understood as R2 = 0.9446 of the regression, a considerable value, indicating that the model is a good fit and can predict the Monthly Income accurately.

文本

描述已自动生成

***Figure 5. Random Forest model results***

As shown in Figure 6, the R2 of train test and test set are similar, which means the model is not overfitted and has high accuracy. Looking at the predictive performance of the model as shown in Figure 6, we can see that it has high accuracy as the dots are distributed around the line.

图形用户界面, 文本, 应用程序

描述已自动生成图形用户界面, 文本, 应用程序

描述已自动生成

图表, 散点图

描述已自动生成图形用户界面, 图表, 散点图

描述已自动生成

***Figure 6. Random Forest model performance***

Based on the random forest regression model that has been constructed, the importance of independent variables can be evaluated from it (Figure 7). The top 3 important independent variables are ‘JobLevel’, ‘TotalWorkingYears’, ‘JobRoleManager’.

"%IncMSE" means increase in mean squared error. By randomly assigning a value to each predictor, if the predictor is more important, the error of the model prediction will increase after its value is randomly replaced. Therefore, the larger the value, the greater the importance of the variable.

"IncNodePurity" means increase in node purity, which is measured by the residual sum of squares and represents the influence of each variable on the heterogeneity of observations on each node of the classification tree, to compare the importance of variables. The larger the value, the more important the variable is.

图形用户界面, 文本, 应用程序

描述已自动生成

图片包含 表格

描述已自动生成

***Figure 7. Importance of the variables***

We can choose between variables based on the cross-validation curve by performing 5-fold cross-validation. The role of the cross-validation method is to try to use different training set/validation set divisions to do multiple sets of different training/validation for the model to deal with the problem of too one-sided test results and insufficient training data. Here we use the training set itself for cross-validation.

The cross-validation curve shows the relationship between the model error and the number of variables used for fitting. The error will first decrease with the increase of the number of variables. The decrease is obvious at the beginning, but when it reaches a certain range, the decrease will no longer change significantly, or even increase.

According to the cross-validation curve (Figure 8), it is suggested that keeping 4-9 important variables can obtain ideal regression results, because the error at this time is the smallest.

图表, 折线图

描述已自动生成

***Figure 8.*** ***Cross-validation curve***

As shown in Figure 7, the top 6 important variables are ‘JobLevel’, ‘TotalWorkingYears’, ‘JobRoleManager’, ‘JobRoleResearch.Director’, etc. We use these top 6 variables to build a new optimized model.

We can see that the 6 variables explained about 93.94% of the total variance, which can be understood as R2 = 0.9394 of the regression, a considerable value, indicating that the model is a good fit and can predict the price accurately (Figure 9).

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

***Figure 9. Optimized Random Forest model***

The prediction accuracy of the new model is shown in Figure 10. We can see that the accuracy is high, and the prediction price and actual price are close to each other.

As shown in Figure 10, the R2 values of train test and test set are similar, which means the model is not overfitted and has high accuracy.

图形用户界面, 文本, 应用程序

描述已自动生成图形用户界面, 文本, 应用程序

描述已自动生成

图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成

***Figure 10. Optimized Random Forest model performance***

**1.3.3 Results:**

The top 6 important variables ‘JobLevel’, ‘TotalWorkingYears’, ‘JobRoleManager’, ‘JobRoleResearch.Director’, ‘YearsAtCompany’, and ‘Age’ are used to build the model to predict Monthly Income. The model accuracy is high with R2 = 0.9394.

We found that ‘JobLevel’, ‘TotalWorkingYears’ have important impact on Monthly Income. If the HR of the company would like to determine the monthly salary of a new employee, these factors should be considered seriously.

**1.4 SVM & Logistic Regression for predicting Attrition**

**1.4.1 SVM**

**1.4.1.1 Method and Reasons:**

SVM method is used to predict the Attrition of the employees. The reasons are shown as below:

SVM is a highly useful technique if we don't know a lot about the data. It can be applied to data with irregular distribution patterns and unknown distribution.

We may handle any complex problem by using the related kernel function, which is a very helpful strategy offered by the SVM. Kernel offers the option of selecting a function that is not always linear and that can take on different shapes depending on the kind of data it uses, making it a non-parametric function.

**1.4.1.2 Analysis Steps:**

To assess the model's predictive power, we must randomly divide the data into a training group and a test group and set the data scale ratio of the two to 7:3.

We used the svm function to build the SVM model using all the remained variables as independent variables. We call tune.svm function to tune the SVM. Then we set up the SVM with the optimal parameters obtained by the tuning function.

We called the confusionMatrix function to evaluate the performance of the optimized model (Figure 11). As can be seen from Figure 10, TP = 15, TN = 365, FN = 48, FP = 2. Accuracy = 0.884, Precision = 0.884, Recall = 0.995, and Specificity = 0.238. The higher the kappa value, the better the training effect. The Kappa value is 0.334, which means the model does not fit very well.

The ROC curve is shown in Figure 12. The two main indicators in the ROC curve are the true rate and the false positive rate. The abscissa is the false positive rate (FPR), and the ordinate is the true rate (TPR). FPR represents the degree of response that the model falsely reports, while TPR represents the degree to which the model predicts the response. The higher the TPR and the lower the FPR (i.e., the steeper the ROC curve), the better the performance of the model. As can be seen from Figure 12, the AUC is 0.616, which is not very close to 1, so we can say that the prediction accuracy of the model is not very high.

表格

描述已自动生成

***Figure 11. Confusion Matrix of SVM model***



图表, 折线图

描述已自动生成

***Figure 12. ROC plot of SVM model***

**1.4.2 Logistic Regression**

**1.2.2.1 Method and Reasons:**

Logistic Regression method is used to predict the Attrition of the employees. The reasons are shown as below:

The predicted parameters (trained weights) provide information on the relative weights of the various features. It also specifies if the relationship is positive or negative. To determine the link between the features, we can utilize logistic regression.

Unlike decision trees or support vector machines, this approach enables models to be quickly modified to incorporate new data. Stochastic gradient descent can be used to update the model.

**1.2.2.2 Analysis Steps:**

Firstly, we used the glm function to build the logistic regression model. From Figure 13, we can see that variables including ‘BusinessTravel’, ’MaritalStatus’, and ’OverTime’, etc. have significant impact on Attrition as their p values are very small. So, we built a new model only including these significant variables (Figure 14).

手机屏幕截图

描述已自动生成表格

描述已自动生成

***Figure 13. GLM model***

表格

描述已自动生成文本

描述已自动生成

***Figure 14. Optimized GLM model***

We called the confusionMatrix function to evaluate the performance of the optimized model (Figure 15). TP = 19, TN = 354, FN = 44, FP = 13. Accuracy = 0.867, Precision = 0.889, Recall = 0.965, and Specificity = 0.302.

文本

中度可信度描述已自动生成

***Figure 15. Confusion Matrix of optimized GLM model***

The ROC curve is shown in Figure 16. The AUC is 0.737, which is not very close to 1, so we can say that the prediction accuracy of the model is very high.

图表, 折线图

描述已自动生成图片包含 徽标

描述已自动生成

***Figure 16. ROC of optimized GLM model***

**1.4.3 Results:**

Logistic regression is better than SVM for predicting Attrition as it has a higher AUC value, simpler model, and similar model accuracy. However, false negatives were high, which would compromise the analysis of this study. If an employee leaves the company and we say he stays, then we lose important records of those who left. We will not be able to analyze the reasons for attrition.

We found that variables including ‘BusinessTravel’, ’MaritalStatus’, ’OverTime’, ’Age’, and ’DistanceFromHome’, etc. are important for Attrition.

**2. Interpretations**

For the first question, we built Random Forest model to predict Monthly Income, and we found that ‘JobLevel’, ‘TotalWorkingYears’, ‘JobRoleManager’, ‘JobRoleResearch.Director’, ‘YearsAtCompany’, and ‘Age’ are the top 6 important variables to predict Monthly Income. Their %IncMSE and IncNodePurity values are higher than other variables, especially the %IncMSE values of ‘JobLevel’ is higher than 40, which indicates that it is very important to Monthly Income. The model accuracy is high with R2 = 0.9394.

If the HR of the company would like to determine the monthly salary of a new employee, ‘JobLevel’, ‘TotalWorkingYears’, ‘JobRoleManager’ should be considered seriously. For example, if the employee has a high job level, long total working years as a manager role, the monthly income must be very high.

For the company, they should use these important variables to determine if the salary of the employees should be adjusted. Monthly income should increase as the total working years and job level increase. Job level is the most important factors of Monthly Income.

For the second question, we built SVM model and logistic regression model to predict Attrition. We found that logistic regression model is better than SVM model as it has a higher AUC value which means it has a better model performance. Also, logistic regression model has a higher model accuracy, and it is much simpler than SVM with fewer independent variables and lower dimensionality. False negative is more damaging for the analysis in this study. If an employee leaves the company and we say he stays, then we lose important records of those who left. We will not be able to analyze the reasons for attrition. Logistic regression has a smaller False negative value than SVM, which indicates that it is more accurate. From logistic regression model, we found that variables including ‘BusinessTravel’, ’MaritalStatus’, ’OverTime’, and ’Age’, etc. are important for Attrition as their p values are very small.

If the HR of the company would like to analyze employees that are likely to leave the company, he should pay more attention to the above important factors. Also, if the company would like to save employees who are going to leave the company, they can lessen the business travel times, working hours and provide apartments near the company for the employees.

**3. Recommendations & Conclusions**

The random forest model for predicting monthly income is very accurate with R2 = 0.9394. ‘JobLevel’, ‘TotalWorkingYears’, ‘JobRoleManager’, ‘JobRoleResearch.Director’, ‘YearsAtCompany’, and ‘Age’ are the top 6 important variables to predict Monthly Income.

Logistic regression model is better than SVM model to predict Attrition as it is more accurate and simpler. Variables including ‘BusinessTravel’, ’MaritalStatus’, ’OverTime’, and ’Age’, etc. are important for Attrition.

We will go further in variables including ‘BusinessTravel’, ’MaritalStatus’, ‘Age’, and ‘NumCompaniesWorked’ as they are very important to attrition. We will figure out the features of employees that are likely to leave.

To better predict employee attrition and find out who is potentially leaving the company, we will set attrition risk metrics for all employees. We can predict the probability of their attrition using the predictive model and set them into different groups.

No risk, if 0 <= fit <= 0.5 Low risk, if 0.5 < fit <= 0.6

Medium risk, if 0.6 < fit <= 0.8 High risk, if 0.8 < fit <= 1

HR can collaborate with the management team to develop an action plan, have discussions about the degree of employee engagement, and pinpoint problems and concerns that need to be addressed right away. Like this, managers of medium-risk employees can keep tabs on behavioral alterations and plan one-on-one or open dialogues. All employees in the low-risk category can have their issues addressed by creating an action plan after understanding them through an open house conversation. There is no immediate need for action for personnel in the no-risk category.

**4. Reference:**

* IBM HR Analytics Employee Attrition & Performance. (2017, March 31). Kaggle. Retrieved October 1, 2022, from https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset?select=WA\_Fn-UseC\_-HR-Employee-Attrition.csv
* Jain, R. A. F.-. R. S. (n.d.). IBM HR Analytics Employee Attrition & Performance. Retrieved October 1, 2022, from <http://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2018FBL/IBM_Attrition_VSS.html>
* O., N. (2021, December 29). Random Forest regression in R: Code and interpretation. HackerNoon. Retrieved October 16, 2022, from https://hackernoon.com/random-forest-regression-in-r-code-and-interpretation
* Conceptual origins, technological advancements, and impacts of using ... (n.d.). Retrieved October 5, 2022, from https://www.researchgate.net/publication/357295200\_Conceptual\_Origins\_Technological\_Advancements\_and\_Impacts\_of\_Using\_Virtual\_Reality\_Technology\_in\_Education

‌