Published in Towards Data Science

Ishan Choudhary Follow

Jun 13, 2020 · 9 min read ★ · Listen

Unsplash: Martek Bjork

IBM HR Attrition Case Study

Predicting if a particular employee is going to leave the organization?

Attrition has always been a major concern in any organization. The time, money and

The IBM HR Attrition Ca ns to identify important factors that might be influential in determining which employee might leave the firm and who may not. This article provides in-depth analysis as well as predictive modelling to understand important factors and make accurate predictions.

Table of Content

Data Preparation and Understanding

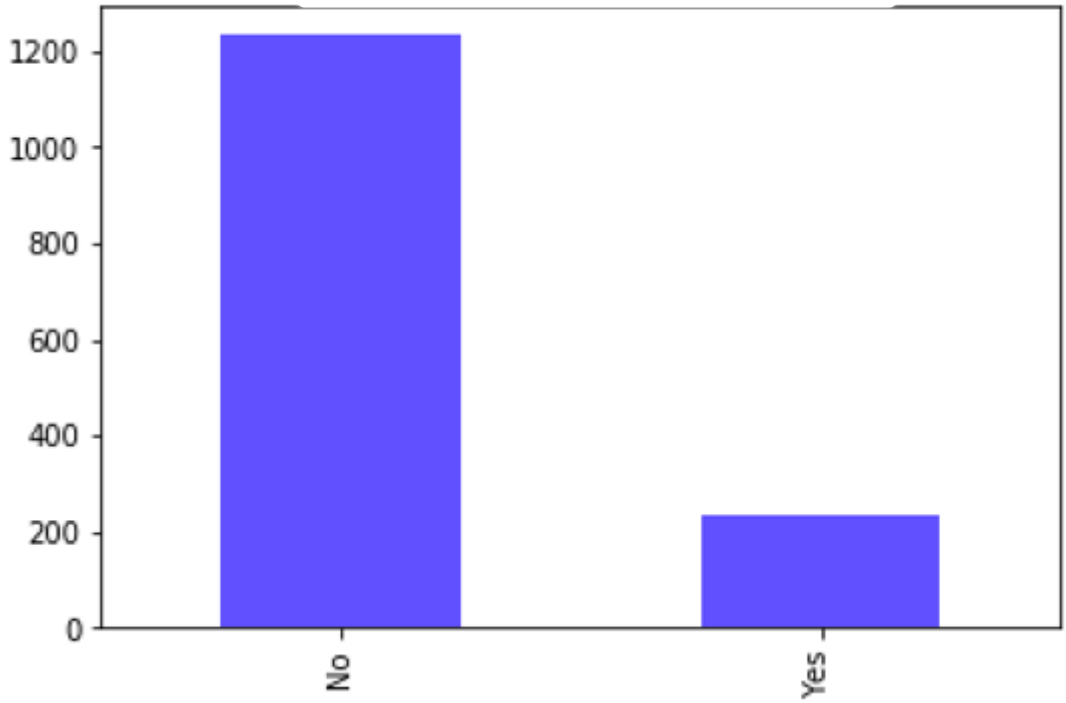
The IBM HR Attrition Case Study can be found on Kaggle. *Python 3.3* is used for analytics and model fitting. The IDE used is *Spyder 3.3.3*.

To properly understand the dataset, let us look at some of its basic features. This includes the shape of the dataset and the type of features/variables present in the data. Furthermore, we will also look at the missing values (if any).

```
In [2]: df = pd.read_csv("Attrition.csv")
   ...: df.shape
Out[2]: (1470, 35)
```

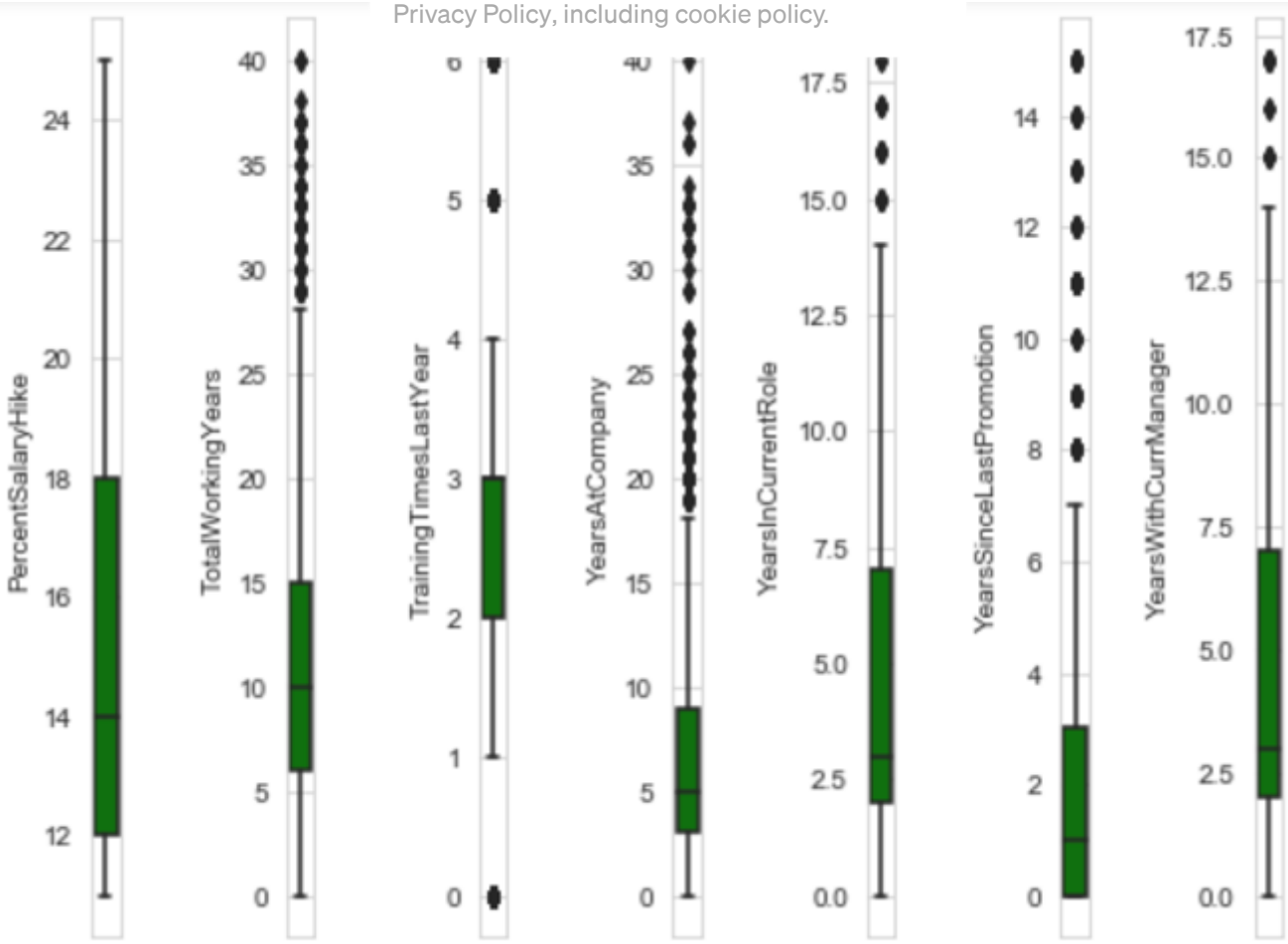
Shape of the Dataset

The *Attrition* dataset had 1470 observations with 35 variables. Out of the 35 variables, there exists one target variable *Attrition* with possible outcomes *Yes* and *No*. The other 34 variables are independent variables but one, that was, *Employee Number* which denotes the employee number or the identification number.

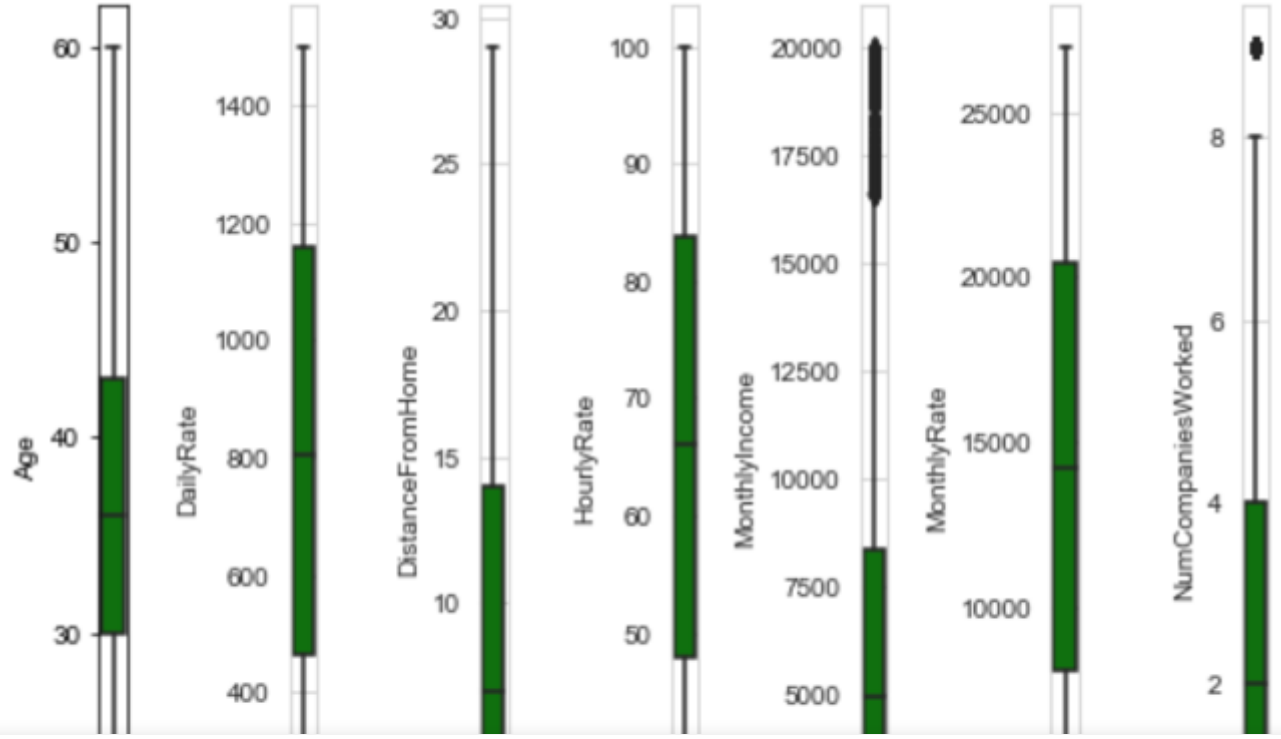



The above graph shows the distribution of the target variable. Out of the total of 1470 observations, 1233 is *No* whereas 167 is *Yes*. We will treat this imbalance after splitting the data into *Training* and *Test Set*.

The dataset has no missing values. Hence, no further treatment is required pertaining to the missing values. Let us use box plot to observe the outliers.



Box Plots (1)



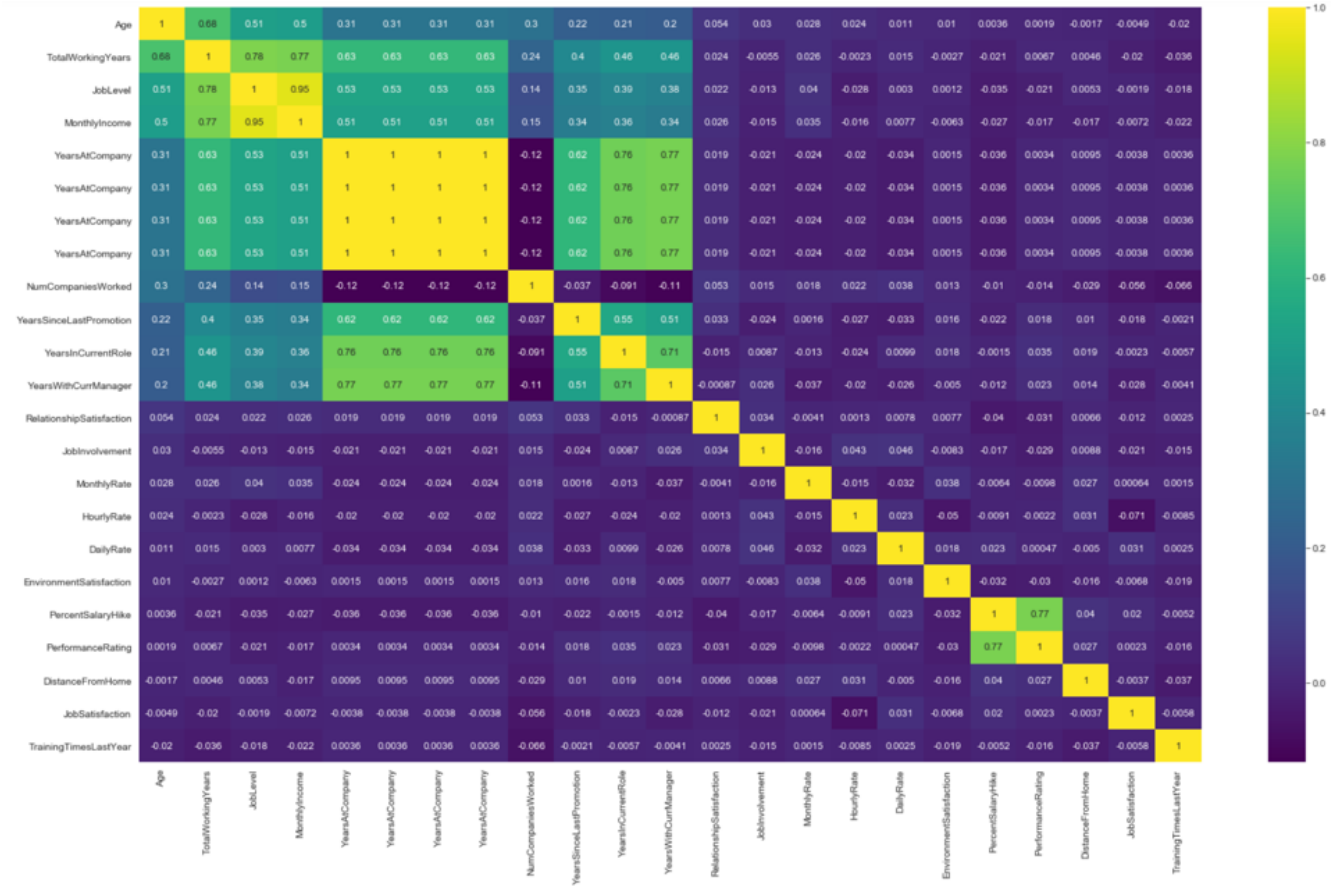
Age, *DailyRate*, *DistanceFromHome*, *HourlyRate*, *MonthlyRate*, *PercentSalaryHike* tend not to have any outliers.

NumCompaniesWorked, *TrainingTimesLastYear*, *YearsWithCurrManager*, *YearsInCurrentRole* have a moderate number of outliers.

MonthlyIncome, *TotalWorkingYears*, *YearsAtCompany*, *YearsSinceLastPromotion* have large number of outliers.

One way to counter this problem is by scaling the variables so as to reduce its effect on the model. The *StandardScaler()* in Python's *Scikit-learn* library can be used for this purpose.

One final step before moving further is to check for multi collinearity. We plot a correlation matrix for this purpose.



Correlation Matrix

multi collinearity is if *co* Based on that we identify the following variables to have a high correlation:

Correlation between *MonthlyIncome* and *JobLevel* is 0.95. This is a very high correlation.

Correlation between *TotalWorkingYears* and *JobLevel* is 0.78 which is also very close to 0.80.

All other variables seem to have a correlation which is less than 0.80.

Having got an understanding of the data along with preliminary analysis, we can now more onto *Feature Engineering*.

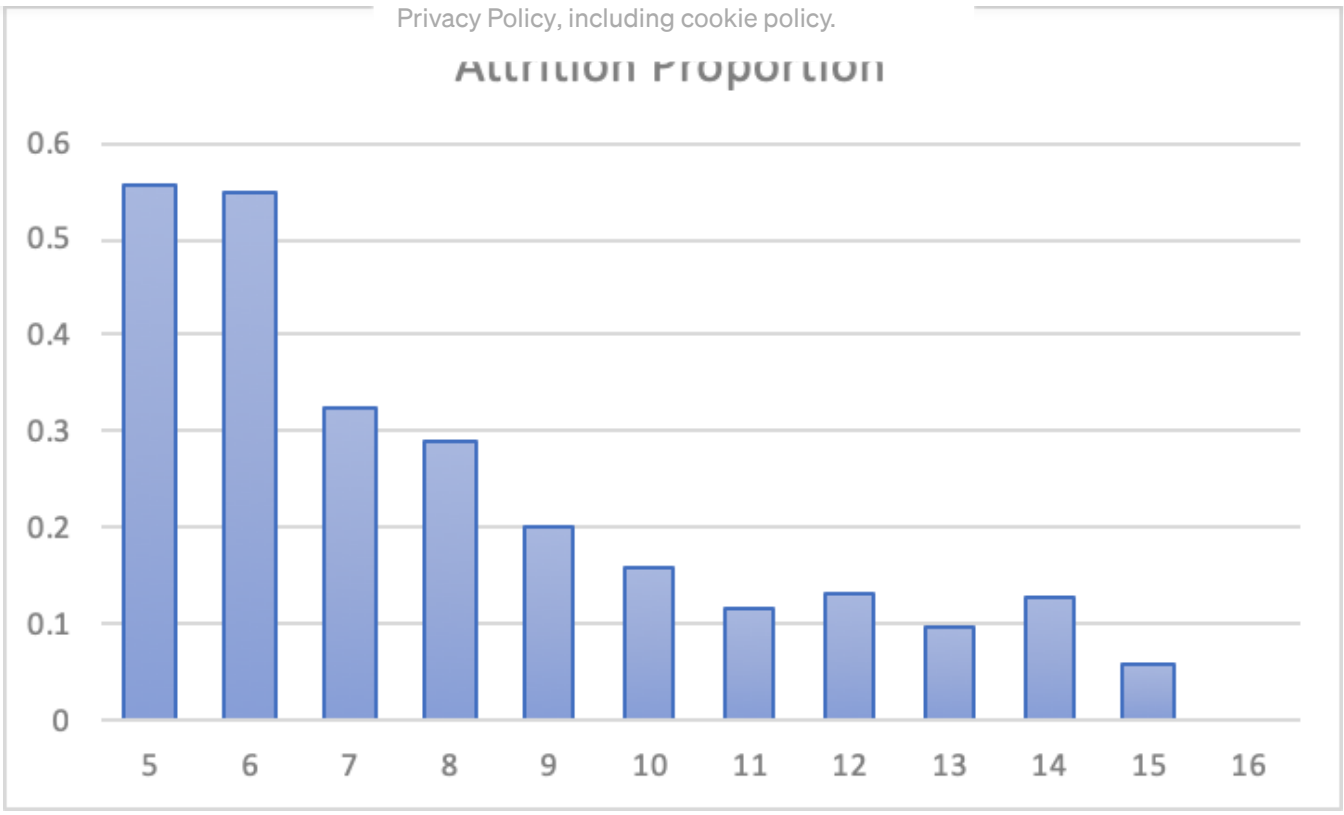
Feature Engineering

Feature engineering refers to a process of selecting and transforming variables when creating a predictive model using machine learning or statistical modeling (such as deep learning, decision trees, or regression). The process involves a combination of data analysis, applying rules of thumb, and judgement.

For the purpose of the model, two features were created from the existing independent variables that existed:

Holistic:

4 factors namely, *EnvironmentSatisfaction*, *JobInvolvement*, *JobSatisfaction, RelationshipSatisfaction* were taken, the sum of which was considered as *Holistic Satisfaction*. The maximum score ’16’ indicated a perfect Holistic Satisfaction whereas ‘4’ was considered as the lowest Holistic Satisfaction. The engineering of this variable was based on the given chart:

Attrition Proportion vs Holistic Satisfaction

The horizontal axis denotes the *Holistic satisfaction* and the vertical axis denotes the proportion attrition corresponding to the given level. It can be clearly concluded that as the grade increases, the proportion of attrition decreases, that is, from 0.56 at grade 5 to 0.00 attrition at grade 16.

BelowAverageIncome

Another feature namely *BelowAverageIncome* was created based on the department the employee was working in, the average income of that department and *PercentSalaryHike* of the employee. If the monthly income of an employee was less than the average income of that department and the percentage salary hike was less than 16, the employee was given a grade 1 else 0, indicating that the employee was most likely to leave.

Feature Selection

Based on the Exploratory Data Analysis and Descriptive Statistical Analysis, we select and deselect certain variables which do not significant contribute to our model. The variables are deselected based on the following:

3. Independence betwe riable
4. Multicollinearity

Variable Type:

The nominal variable is not used in the analysis as it does not provide any input to the model building process. It is however kept so as to identify the employees on whom the study is done.

Invariability in the Data point:

Certain variables do not have any variability in them. Such variables are:

1. *Employee Count*: This is just a count of employee and the value it takes is always 1.
2. *Over18*: This variable describes if an employee is over 18 years of age. It takes the value 'Yes' in all cases.
3. *StandardHours*: The standard number of hours an employee works in a week. Its constant value is 80

Independence between Target variable and Independent variable:

It is important to check the dependence between target variable and independent variable. If there exists insignificant relation between the two, we should not select that independent variable for model building process.

To determine the dependence between Independent variable and the target variable, we use the Chi-Square Test. We set the hypothesis:

H0: There exists no association between the two variables

H1: There is association between the variables.

We reject the null hypothesis if the *p-value* is less than the Level of Significance (a) or if the calculated value is more than the table value. For the purpose of this study, we choose LOS (a) to be at 5%.


Variable		lue
Education	3.074	0.546
Performance Rating	0.000	0.990
Gender	1.117	0.291
DistanceFromHome	38.169	0.095
RelationshipSatisfaction	5.241	0.155

Variables removed from Model building

Multi collinearity:

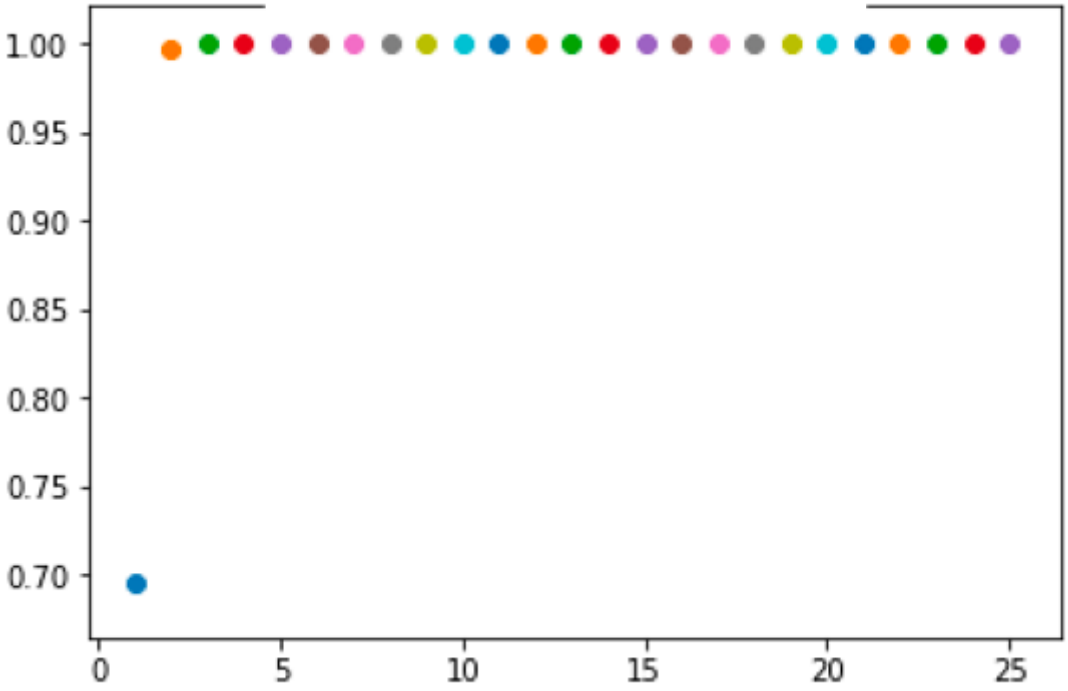
Multi collinearity refers to the strong relationship or correlation between to input variables. There is said to be multi collinearity between two variables if there exist a correlation coefficient of more than 0.80. It is important to remove such variables as this leads to an inflated variance in the model which also increases the error in the model.

Based on our analysis, we remove the following variables:

1. *JobLevel*: 0.95 correlation coefficient with *MonthlyIncome* and 0.78 with *TotalWorkingYears*.
2. *TotalWorkingYears*: 0.77 correlation with *MonthlyIncome*.

Principal Component Analysis

For Feature Extraction, we can also use Principal Component Analysis (PCA). Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing.



Number of Principal Components vs Proportion Variability Explained

Two Principal Components have been created that explains the variability of the 26 variables considered. A loop was run taking the components from 1 to 26 and the explained variability was measured. With only 1 principal component, the explained variability was approximately 68%. This increased to approximately 99% with two principal components and close to 99.99% with 3 principal components. Any further increment increased the percentage explained insignificantly.

Model Fitting

After pre-processing, we split our data into training, validation and test dataset. From a total of 1470 observations, we choose:

1. 80% observation for *Training Dataset.*
2. 14% observation for *Validation Dataset.*
3. 6% observation for *Test Dataset.*

The algorithm chosen is *Support Vector Machine* with *Linear Kernel.* The parameters chosen are given below:

```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3,
gamma='auto_deprecated',
  kernel='linear', max_iter=-1, probability=False,
random_state=42,
  shrinking=True, tol=0.001, verbose=False)
```

The important parameters such as *kernel* is chosen to be '*linear*' in this case. The *random_state* is taken to be 42 as a seed and the model is fitted on 1176 observations.

The next process is to predict the values based on the fitted model for the validation and test set. To calculate the accuracy, precision, recall, true positive and true negative, we create a confusion matrix.

The confusion matrix for Validation set is given below:

$$\begin{pmatrix} 268 & 0 \\ 0 & 28 \end{pmatrix}$$

Confusion Matrix (Validation Set)

The confusion matrix for Test set is given below:

$$\begin{pmatrix} 87 & 0 \\ 0 & 11 \end{pmatrix}$$

Confusion Matrix (Test Set)

We now compute the parameters for accuracy for the 2 datasets:

			score	support
No	1.00	1.00	1.00	168
Yes	1.00	1.00	1.00	28
micro avg	1.00	1.00	1.00	196
macro avg	1.00	1.00	1.00	196
weighted avg	1.00	1.00	1.00	196

	precision	recall	f1-score	support
No	1.00	1.00	1.00	87
Yes	1.00	1.00	1.00	11
micro avg	1.00	1.00	1.00	98
macro avg	1.00	1.00	1.00	98
weighted avg	1.00	1.00	1.00	98

Evaluation Metrics for Validation and Test Set

The first set of information pertains to the Validation set. We get a weighted accuracy of 1.00 with precision as 1.00 and recall as 1.00.

The second set of information pertains to the Test set. We get a weighted accuracy of 1.00 with precision as 1.00 and recall as 1.00.

We now look at the Cohen's Kappa Score. **Cohen's kappa** coefficient (κ) is a statistic which measures inter-rater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as κ takes into account the possibility of the agreement occurring by chance.

```
Kappa score for Training set:  1.0
Kappa score for Validation set:  1.0
Kappa score for Test set:  1.0
```

Cohen's Kappa Score

The Kappa score for Training, Validation and Test set shows that there is absolutely no possibility of prediction by chance. Hence, the results by the model can be completely

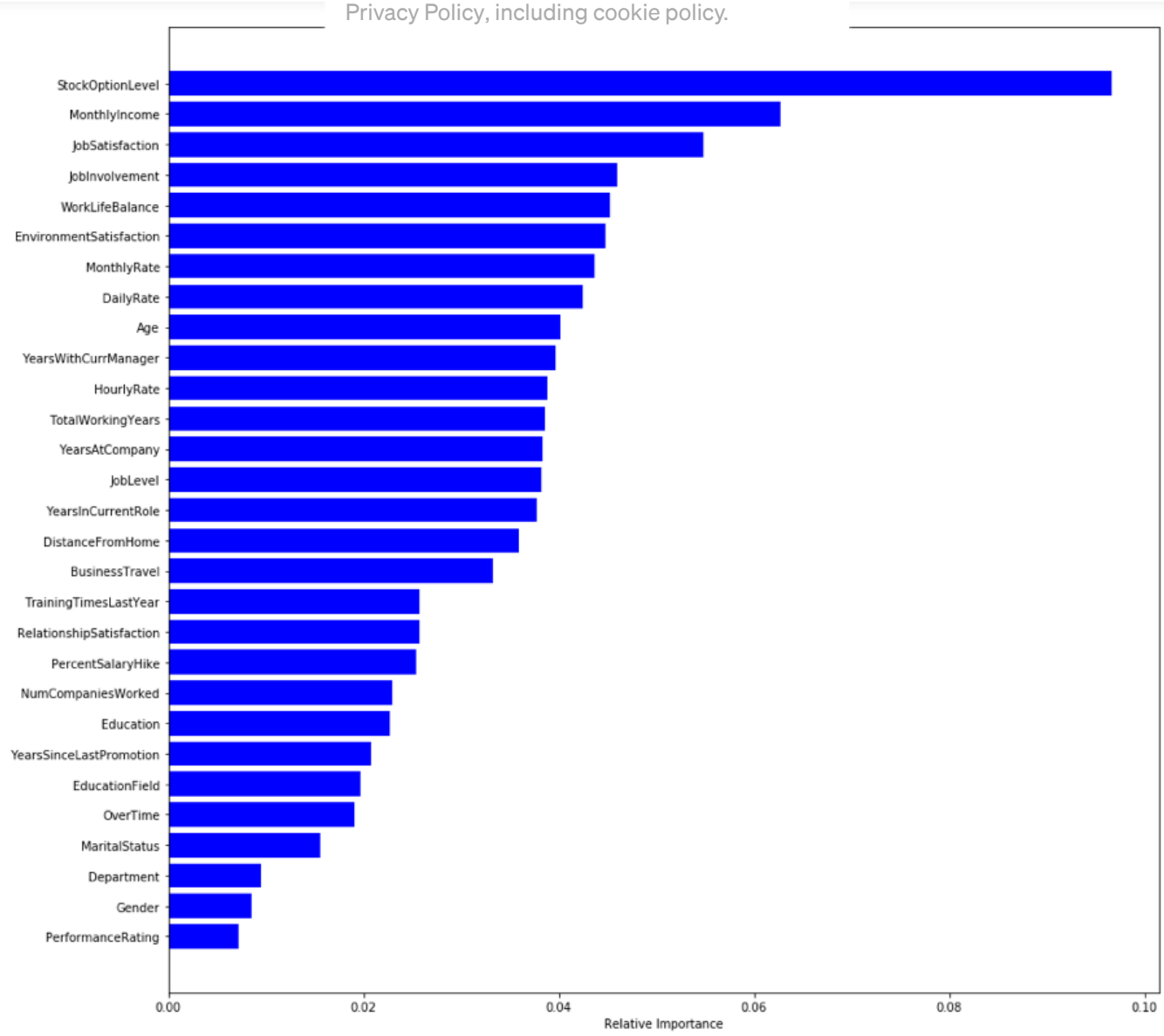
Note: Several other class is explained here. Other classifiers include *Support Vector Machines (RBF Kernel)*, *Random Forest (1000 trees)* and *Logistic Regression*.

Model Comparison

Sr. No	Model Name	Accuracy	Benefits	Trade-offs
1	Support Vector Machine (Kernel: Linear)	100%	Perfect Accuracy	Might succumb to new data points. High Overfitting Time Consuming
2	Support Vector Machine (Kernel: RBF)	79%	Less Time Consuming Moderate Accuracy	Minimal / lower interpretability Overfitting
3	Random Forest	87%	Less Overfitting	Less Flexible Time Consuming

As mentioned, several classifiers were used and the best one was selected. However, it is worth looking at the Benefits and Weaknesses of each classifier. *SVM (Linear Kernel)* gives the highest accuracy (100%) followed by *Random Forest* (87%) and *SVM (RBF Kernel)*. Over fitting can be a concern with the best model as it might succumb to new data points. Also, it takes time in model-fitting. Random Forest, on the other hand can be helpful with new data points in classification.

Recommendations & Conclusions



Based on the above chart, we can conclude that *StockOptionLevel* plays a very important role in deciding the attrition of the employee. Apart from that, *MonthlyIncome, JobSatisfaction, JobInvolvement* also are among the top contributors. On the other hand, factors such as *PerformanceRating, Gender, Department* tend not to contribute as significantly (which was also proven statistically by Chi-Square test).

Based on the overall project, there are certain points that needs to be kept in order to get an optimum output from the model.

The HR Department can focus on the important variables that contribute significantly in determining if an employee is going to leave an organization. Such variables are:

· JobSatisfaction

· JobInvolvement

· WorkLifeBalance

· EnvironmentSatisfaction

Based on the above variables, one can clearly notice a pattern. The employees are more concerned with the materialistic objects that they get directly in hand. Then comes the psychological variables that determines if an employee might leave the organization.

Hence, the HR can focus on such aspects and understand from the viewpoint of the employees. Once that is followed, the project that is called Attrition project can be used as a Retention project. This can immensely help the organization.

Secondly, the model needs to be tuned from time to time as and when new dataset is received. In case any new input variable is introduced, it is important that the information is retrieved for the employees who participated in the initial study.

We hence conclude this project.

Thank you for the read. I sincerely hope you found this article insightful and as always I am open to discussions and constructive feedback.

Drop me a mail at: **icy.algorithms@gmail.com**

You can find me on LinkedIn.