Employee Attrition Control

Dishti Dave(10451780), Dyuti Dave(10453480),
Eman Alofi(10444865), Vaishnavi Gopalakrishnan(10444180)

Abstract

The objective is to present an overview of the machine learning techniques currently in use or in consideration at corporates worldwide. Section I outlines the main reason why corporates should start exploring the use of machine learning techniques. Section II outlines what machine learning is, by comparing a well-known statistical technique (logistic regression) with a (non-statistical) machine learning counterpart (support vector machines). Sections III and IV discuss current research or applications of machine learning techniques. The material presented is the result of a list of machine learning applications on the dataset to predict the status of employees.

I. Introduction

What is machine learning?

In the statistical context, Machine Learning is defined as an application of artificial intelligence where available information is used through algorithms to process or assist the processing of statistical data. While Machine Learning involves concepts of automation, it requires human guidance. Machine Learning involves a high level of generalization in order to get a system that performs well on yet unseen data instances.

Why should corporate companies consider using machine learning?

Machine learning is a relatively new discipline within Computer Science that provides a collection of data analysis techniques. Most statistical techniques follow the paradigm of determining a probabilistic model that best describes observed data among a class of related models. Similarly, most machine learning techniques are designed to find models that best fit data. Therefore, an advantage of machine learning techniques over statistical ones is that the latter require underlying probabilistic models while the former do not. Even though some machine learning techniques use probabilistic models, the classical statistical techniques are most often too stringent for the oncoming Big Data era, because data sources are increasingly complex and multi-faceted. Machine learning might be able to provide a broader class of more flexible alternative analysis methods better suited to modern sources of data. It is imperative for corporate companies to explore the possible use of machine learning techniques to determine whether their future needs might be better met with such techniques than with traditional ones.

II. Classes of Machine Learning algorithms

There are two main classes of machine learning techniques: supervised machine learning and unsupervised machine learning.

Examples of supervised learning:

Logistic regression (statistics) vs Support vector machines (machine learning)

Logistic regression, when used for prediction purposes, is an example of supervised machine learning. In logistic regression, the values of a binary response variable as well as several predictor variables are observed for several observation units. These are called training data in machine learning terminology. The method of maximum likelihood is applied to their joint probability distribution to find the optimal values for the coefficients in this linear function. The model with these optimal coefficient values is called the “fitted

model," and can be used to "predict" the value of the response variable for a new unit for which only the predictor values are known.

Support Vector Machines (SVM) are an example of a non-statistical supervised machine learning technique; it has the same goal as the logistic regression classifier just described: Given training data, find the best fitting SVM model, and then use the fitted SVM model to classify new units.

Examples of unsupervised learning

Principal component analysis (statistics) vs Cluster analysis (machine learning)

The main example of an unsupervised machine learning technique that comes from classical statistics is principal component analysis, which seeks to "summarize" a set of data points in high-dimensional space by finding orthogonal one-dimensional subspaces along which most of the variation in the data points is captured. The term "unsupervised" simply refers to the fact that there is no longer a response variable in the current setting.

Cluster analysis and association analysis are examples of non-statistical unsupervised machine learning techniques. The former seeks to determine inherent grouping structure in given data, whereas the latter seeks to determine co-occurrence patterns of items.

III. Problem Description

Attrition refers to the gradual loss of employees over time. In general, relatively high attrition is problematic for companies. We have a very rich attrition dataset with information ranging from job satisfaction level to basic information about employees. A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork and new hire training are some of the common expenses of losing employees and replacing them.

The main data of interest is "attrition" - STATUS, which indicates whether an employee had any problems in the workplace. Its prediction is very valuable, as it can be used to reduce the number of occurrences and allow the company to better understand the source of conflict. We will try to study the factors that led to employee attrition.
As the value we're trying to predict (attrition- status) is categorical ("yes" or "no"), we're dealing with a classification problem.

Data Preparation & Analysis - Load, Clean and Format

Before creating any Machine Learning models, one must first look at the data with the following questions in mind:

Are there strong relations between the features? Features that are too similar are redundant.
Do I need all the features?
Could the model be simplified somehow?
Could some features be combined, or transformed, to better represent the problem at hand? (e.g. with time variables, sometimes all you need is the hour or day, not the full date).

Many times these questions can only be answered after running a model and testing out some hypotheses. However, looking at some graphs may help this process.

Given infinite resources and time, we could investigate all combinations. However, since there are many columns on this dataset, we're going to choose only a few relations to analyze:

Education Field, Annual Income
Performance Rating, Job Satisfaction

The data is partitioned into two sets: Training(70%), Testing(30%);

- The training set is responsible for initially teaching the model the causal relationship between all information and the attrition probability
- The completed model is applied to the testing set in order to get accurate results on how the model would perform on real-world data

Load data

read.CSV: Used to load the dataset

Clean data

Check for null values in the dataset and replace it with some valid data, if required.

Format data

Change the format of data, if required.

Install & Load required packages

LabelEncoder: converts categorical data into numerical
confusion_matrix: helper method to show the number of correct predictions for 0 and 1 cases
ggplot2: system for creating declarative graphs
corrplot: graphical representation of correlation matrix
dplyr: abstract how the data is stored

gmodels: creates a cross table for model fitting
caret: makes the process consistent, and easy

Load the required packages for each model.

IV. Techniques Used

Data Modelling:

We have our final dataset after cleaning. We now must start modelling- Predicting the Attrition. Most of the time in Regression and Classification problems, you run your model with the available values and check the metrics like accuracy of the model by comparing observed values with true values. If you won't have the true values how would you know that the predictions are correct. Now you will realize how important the training data phase is. We train the model with the training set in a way that it can predict(almost) correct results of the testing set.

kNN

Helps in classifying the data based on how its nearest neighbor is classified. This algorithm normalizes the data by itself.

To substitute the value of K, calculate the square root for the number of observations from the dataset.
K = 9612^0.5 = 98

Include all the training & testing data, the target variable, and the k value in the KNN algorithm.

The prediction shows that for every 30% of the original data the number of employees who are active are more than the terminated.
Total data tested = 2884 (30% of original dataset)
Active employees = 76%
Terminated employees = 24%

Evaluate the accuracy by using the 'gmodels' library that has a 'CrossTable' function. It helps to find out the accuracy. 'CrossTable' returns cross-tabulation of predicted and observed classifications. The table consists of values of positive (TP), true negative (TN), false-negative (FN) and false positive (FP).
Then, apply the values in the following formula to find accuracy;
accuracy <- (tp+tn)/(tp+tn+fn+fp)

kkNN

Performs k-nearest neighbor classification of a test set using a training set. For each row of the test set, the k nearest training set vectors (according to Minkowski distance) are found, and the classification is done via the maximum of summed kernel (k value) densities. In addition even ordinal and continuous variables can be predicted. From the predicted values, find the error rate and accuracy.

Naive Bayes

Naive Bayes is a family of algorithms that all share a common principle: every feature is independent of each other. A Naive Bayes classifier considers each of these features to contribute independently to the prediction, regardless of any correlations between features. This is, of course, almost never true, which means that the applications of this model are very situational. However, it often gives good results, showing better results than other models that do not have their hyper-parameters fine-tuned. It's a pretty simple model.

Decision Tree - CART

Classification And Regression Tree is a simple technique to fit a relationship between numerical variables partitioning the target variable by a range of values of the explanatory variables. This function fits and graphs a cart model with a previous separation of training and testing datasets.

Linear Regression

Regression analysis is a very widely used statistical tool to establish a relationship model between two variables. One of these variables is called a predictor variable whose value is gathered through experiments. The other variable is called the response variable whose value is derived from the predictor variable.

SVM

The next model for fit is a support vector machine(SVM) model. Basically, an SVM constructs a hyperplane or a set of hyperplanes that have the largest distance to the nearest training data points of other classes. Choose a radial kernel with proper gamma and cost values here to optimize the performance of SVM. The SVM model object is a list presenting basic information about the parameters, number of support vectors, etc. The number of support vectors depends on how much slack we allow when training the model. If we allow a large amount of flexibility, we will have many support vectors.

Artificial Neural Network

Neural Networks are a machine learning framework that attempts to mimic the learning pattern of natural biological neural networks. Biological neural networks have interconnected neurons with dendrites that receive inputs, then based on these inputs they produce an output signal through an axon to another neuron. We will try to mimic this process using Artificial Neural Networks (ANN). Fits a single hidden layer ANN model to input data x and output data y.

Random Forest

When working on a classification problem, it's almost always a great idea to start with RandomForestClassifier; not only it's a pretty good classifier for many problems, but it also allows you evaluate what was the impact of each feature on the prediction. This is precious information, as one can build better models by removing undesirable features. There are other ways of figuring out which features are important to our model as well.

C50

Simple to understand, interpret, visualize. Decision trees implicitly perform variable screening or feature selection. Can handle both numerical and categorical data. Can also handle multi-output problems. Decision trees require relatively little effort from users for data preparation. Nonlinear relationships between parameters do not affect tree performance.

k-Means

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as Euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific like our attrition problem and find out clusters related to it.

h-Clustering

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly like each other. Hierarchical clustering starts by treating each observation as a separate cluster.

Then, it repeatedly executes the following two steps:
(1) identify the two clusters that are closest together, and
(2) merge the two most similar clusters. This iterative process continues until all the clusters are merged together.

V. Recommendations

Most of all we found factors which are most important to employees and if they are not fulfilled might lead to Attrition.

VI. Conclusion & Future Work

Conclusion

Machine learning models are as good as the data you feed, and more data would strengthen the model. While some level of attrition is inevitable it should be kept at the minimum possible level using above solution.

Based the most important features to the least important features we can identify what are the main causes for attrition. It helps to understand the key variable that influence the turnover.

We have considered every selective group of features to identify what works best for our dataset and analyzed our model solution

Future Work

New machine learning techniques can be applied to business application and especially predictive analytics.

In this case we can use H2O/LIME to develop and explain sophisticated models that very accurately detect employees that are at risk of turnover.

Or advanced machine learning models and selective features can help improve our predictive analysis. Using ensemble model to breakdown complex structure into critical features that most related to attrition

VII. Related links

GitHub Repository: https://github.com/vaishnavimecit/CS513Spring20_DataMiners