CS513 KDD

Presentation Notes

Neil Gupte

10445674

Slide 1:

Hello , Professor

This is Neil Gupte CWID 10445674.

I have had prior knowledge to data analysis using python but sticking to the course and your teachings I decided using R is the perfect choice for me get some practice and try my skills at actual real world dataset.

I have tried utilizing most of the algorithm we have covered in class as well as tried doing some in depth analysis of a few.

That being said let us begin with the Final Project titled the Data Analysis of employee attrition dataset.

Slide 2:

* Many organizations treat their employees like they’re a dime a dozen. These companies think that even their best workers can easily be replaced, particularly in today’s tricky economy where there’s no shortage of folks looking for work.
* They're wrong.
* Great businesses understand the importance of keeping their employees for a long time.
* Following underscores the importance of employee retention
* **You Can’t Build a Business Without Consistency:**retention is great for ROI.
* **Turnover Crushes the Bottom Line:** According to the [Center for American Progress](https://www.americanprogress.org/issues/labor/report/2012/11/16/44464/there-are-significant-business-costs-to-replacing-employees/), it costs roughly 20% of an employee’s salary to replace that individual. So when your employees are constantly leaving for greener pastures, you’re forced to spend the equivalent of over two month’s worth of salary just to find someone to replace them.
* **You Lose Talent and Ideas:**
* **It’s Difficult to Establish Camaraderie With a Transient Workforce**
* **Customers Notice When Names Change:**Your clients have been interacting with employees for some time. When the people they’re used to dealing with leave, it can put a bad taste in their mouth, leaving your future working relationship with that organization in doubt.
* **Job Seekers Notice Too:** Thanks to sites like [Glassdoor](https://www.glassdoor.com/index.htm), it’s easier than ever to get a glimpse into what it’s like to work at an organization — even if you just found out that organization exists in the first place. Almost 50% of candidates use Glassdoor during their job search, according to [Software Advice](http://new-talent-times.softwareadvice.com/how-job-seekers-use-glassdoor-0114/).
* **Constantly Training New Employees Is a Waste of Resources:** Those managers you hired? You almost certainly didn’t hire them to spend their time training new hires over and over and over again. You hired them to help employees develop and to grow your business.
* **Your Competitors Could Benefit Directly :** When you lose a talented employee, maybe that person decided to go back to school, move across the country, or switch industries altogether.Or maybe that individual sought employment at one of your competitors — the one that seems to be most respected and offers its employees the best perks. Worse yet, maybe that organization went on the offensive and approached your employee first.

Slide 3:

This is the brief snapshot to what we are given and what we will be dealing with to gain some credible information from.

Slide 4:

**Exploratory Data Analysis (EDA)**, also known as Data Exploration, is a step in the Data Analysis Process, where a number of techniques are used to better understand the dataset being used.

Using summary function available in R we get a gist of all necessary important stats about the data.

While doing it we find out there are 5394 missing values in TERMINATION\_YEAR column. Which I decided to replace with a zero. There are various algorithms which we will be using further which do not allow missing values in the data.

R provides with library called **ggplot2 ,** which provides aesthetically pleasing functions to visualize the data.

Slide 5:

A **frequency distribution** is an overview of all distinct values in some variable and the number of times they occur.

We use the function **geom\_bar()** in ggplot to draw the frequency bar charts.

Thus we can analyze the distribution of the status in each column.

In overall dataset there are more active people data than the terminated people data.

From the graph on right we get to see that the number of female employees in the company is more. And the percentage of termination is more in females than in males.

Slide 6:

Continuing with the frequency distribution graphs,

The graph of the left shows the ethnicity distribution data ,where we find the number White employees in the company is lot more and next comes the Asians.

The graph on the right shows the marital status distribution, where we find the maximum people are single or married, while the number of divorced people are very less.

Interesting Fact is the termination percentage seems to be more in singles than married people.

Slide 7:

A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are.

We use **geom\_boxplot()** function in ggplot2 to plot these.

The graph on the left shows the box plot wrt the Annual rate, where we see lot of outliers ,those are the people earning way more than the average positions .

We see that the median of people who are terminated have lower income range than the active group. Thus people with lower income are more susceptible to termination.

The graph on the right shows the boxplot wrt the Age. Where we find that average group of age is between 27-53 yrs and we also find the people who are terminated have slightly older starting age than active group.

Slide 8:

Scatter Plot is a graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present.

We use **geom\_point()** function in ggplot2 to plot these.

From this graph we see to it that the whites have lot of representation in all salary ranges. Also mostly the person heading this company is of a black race .

We see termination rates are very high at lower income ranges.

Slide 9:

This is the scatter plot of hourly rate wrt sex, We get to see that the data has more number of women ,and probably there are lot of women at higher business positions in the company.

Slide 10:

This graph shows the scatter plot of age wrt the marital status. Thus we get to see that the ages of people who are divorced fall into slightly younger age group of 18-40.

Slide 11:

A **histogram** is a graphical display of data using bars of different heights. In a **histogram**, each bar groups numbers into ranges.

We use **geom\_histogram()** function in **ggplot2** to plot these.

The graph shows us the that there are very few people who are 19 years of age in the company while the 37 years has the highest number of people. Also we see a pattern where the termination rates is higher at in the younger age group from 18-35 and older age group of 55-64.

Slide 12:

This is another frequency distribution barchart for job satisfaction, here we see that maximum number of people have chosen 1,2. The highest termination percentage is found in group 5 wrt the number of people considered.

Slide 13:

This graph shows the frequency distribution bar chart for performance rating. We see an interesting pattern here where the termination percentage is higher in group 2 and 4. While group 3 has highest number of people.

Slide 14:

A classification algorithm, in general, is a function that weighs the input features so that the output separates one class into positive values and the other into negative values.

These following are the algorithms used to classify our dataset . We go on building models using each algorithm to figure which gives the best results .

Slide 15:

* KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique we generally look at 3 important aspects:
* 1. Ease to interpret output
* 2. Calculation time
* 3. Predictive Power
* **KNN** is a **non-parametric.** When we say a technique is **non-parametric** , it means that it does not make any assumptions on the underlying data distribution. In other words, the model structure is determined from the data.
* After doing the min max normalization , we created unweighted knn model using the **kknn** library provided in R and parameter of kernel=“rectangular”.
* The following is the confusion matrix we observe from the model and we get accuracy of 61.06103%.

Slide 16:

Next we do weighted knn using the parameter of kernel =“triangular”. The following is the confusion matrix we observe.

The weighted knn produces accuracy of 60.26352% which is slightly lower than unweighted knn.

Slide 17:

A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task. The crux of the classifier is based on the Bayes theorem.

We use the **naiveBayes()** function in R to build the model.

We get an accuracy of 61.1997% from the model which is similar to knn.

Slide 18:

A Decision Tree is a simple representation for classifying examples. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter.

We use the **rpart** library to build the model and **rattle**  to visualize the model.

From the dtree plot we can see that the algorithm chooses the following parameters as tree roots. We get to see that if the prev\_5 year rating is greater than or equal to 0.5 then he has 79.21 % probability of being active. It means that there is a pattern where employees who have worked of more than 5 years seem to be loyal to the company. Thus we can read the whole tree and obtain needed information.

Slide 19:

The following is the confusion matrix we observe.

We get an accuracy of 62.55201% from the model.

Slide 20:

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning). Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction .

We have the parameter importance graph where in :

Mean Decrease in Accuracy the number or proportion of observations that are incorrectly classified by removing the feature (or values from the feature) in question from the model.

GINI importance measures the average gain of purity by splits of a given variable. If the variable is useful, it tends to split mixed labeled nodes into pure single class nodes. Splitting by a permuted variables tend neither to increase nor decrease node purities.

Thus we get the important parameters to consider in our classifier using this which we plot next.

Slide 21:

We see the important parameters with the most important at the top . These factor will have more weightage in the model.

Slide 22:

We use **randomForest()** function available inside the **randomForest** library to build the model.

We get an accuracy of 63.34951% from the model.

Slide 23:

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.

Here we use the **neuralnet()** function from **neuralnet** library to build the model.

We build the model with 5 neurons in the hidden layer .

Slide 24:

We get an accuracy of 56.90014% from the model.

Slide 25:

After building all the models and comparing their accuracies we found out that Random Forest Classifier produces the best results for the given dataset.

We get an accuracy of 63.34951%. Random Forest is especially good because of the following reasons:

1. **Random Forest is Versatile: I**t can handle binary features, categorical features, and numerical features. There is very little pre-processing that needs to be done. The data does not need to be rescaled or transformed.

**2)Parallelizable**: They are parallelizable, meaning that we can split the process to multiple machines to run. This results in faster computation time.

**3) Great with High dimensionality**

**4) Quick Prediction/Training Speed**

**5) Robust to Outliers and Non-linear Data**

**6) Low Bias, Moderate Variance**

**There is still a lot of room to increase the accuracy of the model, hyper tuning some of the parameters can produce better results than what we have acquired.**