**Final Project**

Chun Ning Yip

Webster University

CSDA 6010

Dr. Ali Ovlia

December 12st, 2022

Table of Contents

**Case Study 1 （Predicting Taxi Cancellations in Bangalore, India）3**

Abstract and data set4

Attribute interpretation6

Dimensional Reduction7

Data exploration13

Data partition18

Model 1 (C.50 classification tree)18

Model 2 (rpart classification tree)19

Model 3 (logistic regression)20

Results and conclusion22

**Case Study 2 (Predicting Mortgage Payback in RMBS, U.S**.)2**3**

Abstract and data set24

Attribute interpretation26

Dimensional Reduction29

Data exploration31

Data partition38

Model 1 (K-nn)39

Model 2 (C.50 classification tree)41

Model 3 (Random Forest)42

Results and conclusion43

**Case Study 3 (Human Resource Descriptive Analytics**)45

Abstract 46

Attribute interpretation46

Data exploration48

Hypothesis 153

Hypothesis 254

Hypothesis 355

Data partition58

Models Result 59

**Case Study 1**

**Predicting Taxi Cancellations in Bangalore, India**

Chun Ning Yip

Webster University

CSDA 6010

Dr. Ali Ovlia

September 14th , 2022

**Abstract**

The Bangalore- based taxi company has noticed a systematic problem with their drivers in 2013. They have a very high cancelation rate by driver comparing to others which leads to a problem, lost in customer. Customer don’t want to schedule rides with Bangalore- based taxi company when they got cancelled on necessary rides, some of the customers’ rides got cancelled even having scheduled drivers in advance.

Then, they have deployed an online booking system to try reducing the cancellation rate, it didn’t work at all. This becomes a bigger problem when they have a competitor join, Uber. This case study is to help this taxi company finding out the reason of cancellation and how to prevent it by data exploration and making predictive models which predict whether a ride will be cancelled by the driver.

**Dataset**

These data are collected from year 2011 to 2013, so it will not be cooperated with business, the result won’t be accurate for now 2022. It contains 19 attributes and 10000 records. Attributes included user ID, Vehicle ID, Travel Type ID, etc., will be the predictor variables. The outcome variable is Car\_ cancellation.

|  |  |
| --- | --- |
| **Row ID** | This is the unique identifier of this dataset. It is a number identifying each record. |
| **User ID** | This is the identifier of each customer. There are many duplicates in this subset, meaning there are any customers who have called multiple rides. |
| **Vehicle Model ID** | This is an ID that represents the type of vehicle driven for each ride. |
| **Travel Type ID** | This is an ID that represents the type of travel (1= long distance, 2= point to point, 3= hourly rental). |
| **Package ID** | This is an ID that represents the type of travel package, with the following descriptions: 1=4hrs & 40kms, 2=8hrs & 80kms, 3=6hrs & 60kms, 4= 10hrs & 100kms, 5=5hrs & 50kms, 6=3hrs & 30kms, 7=12hrs & 120kms. |
| **From Area** | This is an identifier of the starting area. Available only for point-to-point travel. |
| **To Area** | This is an identifier of the ending area. Available only for point-to-point travel. |
| **From City ID** | Unique identifier of the starting city. |
| **To City ID** | Unique identifier of the ending city. |
| **From Date** | Date and time of the requested trip start. |
| **To Date** | Time stamp of trip end. |
| **Online Booking** | A binary (0,1) variable representing whether the booking was made online or not. 0 represents no, 1 represents yes. |
| **Mobile Site Booking** | A binary (0,1) variable representing whether the booking was made on their mobile site or not. 0 represents no, 1 represents yes. |
| **Booking Created** | Date and time of booking created. |
| **From Lat** | The latitude of the start area. |
| **From Long** | The longitude of the start area. |
| **To Lat** | The latitude of the end area. |
| **To Long** | The longitude of the end area. |
| **Car Cancellation** | The target variable. A binary (0,1) variable representing whether or not the ride was cancelled. 0 means no, 1 means yes. |

**Attributes interpretation**

Text, letter

Description automatically generated

Figure 1

In figure 1, id variables are presented by integers, date and booking\_ created variables are in characters, booking and cancellation variables are integers with 1 and 0, lastly latitude and longitude variable are presented by double.

Chart, histogram

Description automatically generated

Figure 2

In figure 2, it shows the numbers of driver cancel and not cancel the ride. The 1 and 0 in y-axis is presenting cancelled and not cancelled respectively. The number of cancellations captured in this data set are just an extremely small part of the total.

**Dimensional Reduction**

Distance between each ride can be calculated by values of Latitude and longitude. The formular used are shown below.

L1 = from latitude; L2 = to latitude; Lg1 = from longitude; Lg2 = to longitude; rad =

R = 6378.145

D = R\*C

After we calculated the distance, the latitude and longitude attributes will be removed, and the distance will be rounded up to nearest whole number. Then row\_ id and user\_ id will also be removed, there are multiple user\_ id repeated in the dataset which may give us information, but these information from these two attributes didn’t help us to find the reason why the driver cancel the ride.

Table

Description automatically generated with medium confidence

Figure 3

In figure 3, it shows the number of NAs and Zeros inside the four variables, “*from\_area\_id, to\_area\_id, from\_city\_id and to\_city\_id*”. Three of these variables have more than 2000 missing values, which can’t provide us accurate information and distance has already been calculated, so “*from\_area\_id, to\_area\_id, from\_city\_id and to\_city\_id”* will be removed.

Text

Description automatically generated

Figure 4

In figure 4, it shows that there are 339 rides in travel type 1, 7909 rides in travel type 2 and 1752 rides in travel type 3. Rides in travel type 1 and 2 have 339 and 7909 NAs in travel type id, these results proves that package id is only for travel type 3. So, it will be removed.

Text

Description automatically generated with low confidence

Chart

Description automatically generated with low confidence

Figure 5

Those variables containing date information in this data set are stored as a character, so we will need to transform them into date. In figure 5, it shows that there are 4178 NAs inside to\_ date variables, the number of NAs are almost half of the dataset, so it will be removed.



After the class of both variables become date, how many days in advance did customer book the ride can be found, and the results will replace the booking\_created variable. Weekdays will be extracted from date in from\_date variable and replaced it. These processes will give us more in-deep information then just date.



Text

Description automatically generated

Figure 6

In figure 6, it shows that there are still 2091 NAs inside distance variable, and all of them are from travel type 1 and 3, instead of removing them I decided to put Zeros inside. Then I decided to classify the distance into short, medium, and long, also classify the weekdays into weekend or not.

Chart, bar chart, histogram

Description automatically generated

Figure 7

Weekends will be Friday to Sunday and the rest will be business days, the classification of distance will be based on the density plot show in figure 7. Zeros are removed from the plot, most of the distance of rides are in between 17975 and 18000, so lower than 17975 will be short and higher than 18000 will be long, the rest will be medium.

After the dimensional reduction and attributes interpretation, there are 8 variables left and 10000 records remain as shown in figure 8.

Text

Description automatically generated with medium confidence

Figure 8

|  |  |
| --- | --- |
| **Vehicle Model ID** | This is an ID that represents the type of vehicle driven for each ride. |
| **Travel Type ID** | This is an ID that represents the type of travel (1= long distance, 2= point to point, 3= hourly rental). |
| **From\_ date** | This represent if the ride is on weekend or business day |
| **Online Booking** | A binary (0,1) variable representing whether the booking was made online or not. 0 represents no, 1 represents yes. |
| **Mobile Site Booking** | A binary (0,1) variable representing whether the booking was made on their mobile site or not. 0 represents no, 1 represents yes. |
| **Booking\_created** | The number of days that the customer booked the trip in advance |
| **Car Cancellation** | The target variable. A binary (0,1) variable representing whether or not the ride was cancelled. 0 means no, 1 means yes. |
| **Distance** | The total travel distance round up into 1.00km |

**Data exploration**

Chart, histogram

Description automatically generated

Figure 9

Chart, histogram

Description automatically generated

Figure 10

Chart, treemap chart

Description automatically generated

Figure 11

In figure 9 to 11, these plots show the number of rides in each variables type. Most of the customer booked rides with vehicle model 12, usually created on the same day and in point-to-point travel. Driver may choose to pick up customer near them to save the gas money, so they canceled the rides which is father away when they received the booking from customer. Rides that most of the customers booked is neither from online nor mobile site. The number of rides customers booked on business day are slightly higher than weekends. Most of the rides has a medium distance.

Chart, histogram

Description automatically generated

Figure 12

In figure 12, vehicle model ids are on the x-axis and each of the color represent the travel type id. Focus on vehicle model id 12, most of the rides are from point-to-point travel which is travel type id 2 in green color.

Chart

Description automatically generated

Figure 13

In figure 13, it shows a bar chart with the distance vs travel type, medium and long distance both have only a green bar, it means that point-to-point travel is very effective in those cancellation.

Before creating partition and models, we need to balance the dataset, cancellation of rides is nearly 20% of the whole data set which show in Figure 2. After I do an oversampling to the data, there are 9093 cancellation and 9257 not cancelled which show in Figure 14

A picture containing text

Description automatically generated

Figure 14

Next using logistic regression to check if there is any unimportant variable. The result is shown in figure 15, every value is important, and no variable needs to be removed.

Text

Description automatically generated

Figure 15

**Data partition**

60% of the data set will be training and the rest 40% will be the validation.

A picture containing letter

Description automatically generated

**Model 1 (C.50 classification tree)**

Setting the positive class as 0 and trails = 5. It has a accuracy with 0.737, sensitivity with 0.731 and specificity with 0.743.

**Text

Description automatically generated**

**Model 2 (rpart classification tree)**

Diagram

Description automatically generated

Setting the minsplit = 10 and minbucket=3. Cut off-point as 0.6 and positive class as 0, It has an accuracy with 0.7002, sensitivity with 0.7334 and specificity with 0.6665.

**A screenshot of a computer

Description automatically generated with medium confidence**

**Model 3 (logistic regression)**

**Table

Description automatically generated**

**Chart

Description automatically generated**

**A screenshot of a computer

Description automatically generated with medium confidence**

From the ROC graph, I picked the 17th point as the cutoff point and it has an accuracy 0.6708, sensitivity 0.5918 and specificity 0.7512.

**Results**

I repeated the model on unbalanced dataset and here is the result table comparing the three models’ accuracy, sensitivity, and specificity between balanced and non-balanced data set.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Balanced Dataset | | | Unbalanced Dataset | | |
|  | C.50 | rpart | Logistic regression | C.50 | rpart | Logistic regression |
| Accuracy | 0.737 | 0.7002 | 0.6708 | 0.5044 | 0.5044 | 0.6712 |
| Sensitivity | 0.731 | 0.7334 | 0.5918 | 1 | 1 | 0.5945 |
| Specificity | 0.7432 | 0.6665 | 0.7512 | 0 | 0 | 0.7492 |

From the result, the accuracy of C.50 and rpart are above 0.7 and their sensitivity are close together, but the specificity of C.50 is higher than rpart. C.50 and logistic regression has a better detection on cancellation by looking at the specificity. Next, if the model is created from unbalanced dataset, they couldn’t detect the cancellation and gives 0 in specificity. Still logistic regression for balanced and unbalanced are similar. Using logistic regression as a standard, I will say C.50 is the best model.

**Case Study 2**

**Predicting Mortgage Payback in RMBS, U.S**

Chun Ning Yip

Webster University

CSDA 6010

Dr. Ali Ovlia

October 9th, 2022

**Abstract**

Loan company makes profit from the borrowers when they can pay off their loan with interest. So, investors would like to know the risk they will take when they lend the money. In this case study we are using data provided by international financial research, they provide institutions with advisory services in the areas of financial risk management and prudential supervision. Models will be created to classify if the borrower going to pay off their loan in the end.

**Dataset**

This data covers 622489 loan records between 120 periods of time. They are all random selection of portfolios underlying U.S. residential mortgages. There are totally 22 predictors variables, and the outcome variable is status\_ time.

|  |  |
| --- | --- |
| **id** | Borrower ID |
| **time** | Time stamp of observation |
| **orig\_ time** | Time stamp for origination |
| **first\_ time** | Time stamp for first observation |
| **mat\_ time** | Time stamp for maturity |
| **balance\_ time** | Outstanding balance at observation time |
| **LTV\_ time** | Loan-to-value ratio at observation time, in % |
| **interest\_ rate\_ time** | Interest rate at observation time, in % |
| **hpi\_ time** | House price index at observation time, base year = 10 |
| **gdp\_ time** | Gross domestic product (GDP) growth at observation time, in % |
| **uer\_ time** | Unemployment rate at observation time, in % |
| **REtype\_ CO\_ orig\_ time** | Real estate type condominium = 1, otherwise = 0 |
| **REtype\_ PU\_ orig\_ time** | Real estate type planned urban development = 1, otherwise = 0 |
| **REtype\_ SF\_ orig\_ time** | Single-family home = 1, otherwise = 0 |
| **investor\_ orig\_ time** | Investor borrower = 1, otherwise = 0 |
| **balance\_ orig\_ time** | Outstanding balance at origination time |
| **FICO\_ orig\_ time** | FICO score at origination time, in % |
| **LTV\_ orig\_ time** | Loan-to-value ratio at origination time, in % |
| **Interest\_ Rate\_ orig\_ time** | Interest rate at origination time, in % |
| **hpi\_ orig\_ time** | House price index at origination time, base year = 100 |
| **default\_ time** | Default observation at observation time |
| **payoff\_ time** | Payoff observation at observation time |
| **status\_time** | Default (1), payoff (2), and nondefault/nonpayoff (0) observation at observation time |

**Attribute interpretation**

**A picture containing text

Description automatically generated**

Figure 16

In figure 16, there is a relationship between the id and time, when id repeated, the time will plus one each time. Looking at Figure 17 below, status time only shows 1 or 2 in the latest time of each id records. Since, this data only contains records in between the 120 time periods, some of them still not finished their payment and have 0 in status time as their last record. This also explains the huge percentage of Zeros appeared in status time which show in Figure 18.

A screenshot of a computer

Description automatically generated with low confidence

Figure 17

Chart

Description automatically generated

Figure 18

Figure 19 is a plot showing the balance of id 23195 changing according to time. The balance decreased rapidly after the 30 periods.

Chart

Description automatically generated

Figure 19

In Figure 20, the last few rows’ records contain borrower id 50000, this means that there are only 50000 unique borrowers in these 622489 records.

A screenshot of a computer

Description automatically generated with low confidence

Figure 20

Deal to the special of this data structure, I will extract the last record of each id based on the time then dropped the records with zeros in status time to create the training set and validation set. So, I will have the records which only contains default or payoff. Data exploration will be compared the latest and the first record of each borrower.

**Dimensional Reduction**

By deducting maturity time with origination time, we can calculate how much time do the borrower have to pay off their loan. I want to know if there is a relationship between total time and whether they can payoff. Then orig\_time and first\_time will be removed. No further information is needed to be extract.

Default time and pay off time variables will also be removed because status\_ time has already contained the result. Next, extracting the last record of each id. Separate variables with original time, so we can have the first record of each id and the last record of each id. Figure 21 shows the original dataset, figure 22 shows the dataset after splitting.

Text, letter

Description automatically generated

Figure 21

A screenshot of a computer

Description automatically generated with low confidenceA screenshot of a computer

Description automatically generated with low confidence

Figure 22

Time and id variable are used to splitting the 50000 unique records for each id from the original dataset. After splitting the dataset, they become useless so that they will be removed.

Graphical user interface, text, application

Description automatically generated

Figure 23

There is one last problem, there are still 18 NAs inside the LTV\_ time, based on the explanation it is Loan-to-value ratio, since there is no variable to predict the value, and so there is no way to calculate these NAs, they will be removed.

**Data exploration**

Chart, histogram

Description automatically generated

Figure 24

Figure 24 shows the relationship between maturity time and payoff. 0 is unknown which is in red, 1 is default which is in green and 2 is pay-off which is in blue. Having a longer maturity time won’t have a higher percentage for borrower to pay-off their loan. Most of them pay-off their loan when they have only 120 period of times.

Chart, histogram

Description automatically generated

Figure 25

Figure 25 shows the relationship between balance and pay-off. 0 is unknown which is in red, 1 is default which is in green and 2 is pay-off which is in blue. A bigger blue lines area is on the left of the graph which means that borrowers tend to pay-off their loan when they have lower balance.

Chart, bar chart, histogram

Description automatically generated

Figure 26

Figure 26 shows the relationship between interest rates and pay-off. 0 is unknown which is in red, 1 is default which is in green and 2 is pay-off which is in blue. Some of the outliers like 0 and 38 interest rates are removed. Most of the borrower has an interest rate in between 5 to 10, at around 5 interest rate, percentage of borrowers pay-off their loan is the highest.

Chart, waterfall chart

Description automatically generated

Figure 27

In figure 27, it shows the relationship of three different real estate type with pay-off. On the left side is condominium, middle is planned urban development and the right is single-family home. 0 and 1 on the x-axis represent if it is not the type or is the type of real estate. Status of each borrower shows by the color, 0 is unknown which is in red, 1 is default which is in green and 2 is pay-off which is in blue. Most of the borrower come from single family. The percentage of different pay-off status between these three-estate type are similar to others.

Next, will be comparing the balance, LTV, interest rate and hpi between the last record and the first record, trying to find out a relationship between the difference and the last status. The difference will be calculated by deducting the latest record with the origination time record.

Chart, histogram

Description automatically generated

Figure 28

Figure 28 showing the difference in balance, most of the balance decreased at the end of the observation period. 0 is unknown which is in red, 1 is default which is in green and 2 is pay-off which is in blue. The blue line area appears under 0, means that the borrower has pay-off their loan and so the difference will near zeros.

Chart, histogram

Description automatically generated

Figure 29

Figure 29 showing the difference in LTV, most of them lies between -100 to 50. Similar with the results in figure 28, most of the records lies around 0, and they have a higher amount of borrower pay-off the loan.

Chart

Description automatically generated

Figure 30

Figure 30 shows the difference between interest rate, there is a huge box on zero. Also, the borrower who pay-off the loan tends to have higher interest rate compare with the rate in origination time.

Chart, histogram

Description automatically generated

Figure 31

Figure 30 shows the difference between hpi, the borrower who default at the end tends to have increased in hpi, the borrower who pay-off at the end tends to have decreased in hpi. In conclusion of these comparisons, most of the borrower who pay-off at the end will have a steadier change in these four values.

**Data Partition**

I will do a logistic regression first to find out the important variables, since glm() in r only takes outcome variables value in between 1 and 0, I set the status time value 2 into 1 and 1 into 0. So, in status time variable, 1 means pay-off and 0 means default.

A screenshot of a computer

Description automatically generated with low confidence

Figure 32

In figure 31, the three Real estate types’ p-value is 0.7, 0.9 and 0.03 which is not significant, so they will be removed.

The training and validation will be 60% and 40% of the data set which contains the latest record of each unique id. Records with 0 in status\_ time will be removed, because I only want to predict if borrower will pay-off or not. Lastly, I will do a predict with the best model on the origination time records and see how well it will be, variable name in it needs to be changed to do prediction.

**Model 1 (k-nn)**

**Chart, scatter chart

Description automatically generated**

Figure 33

In figure 32, it shows the result of knn tuning, we can see the even numbers of k will have a higher accuracy then the next odd numbers of k.

**A screenshot of a computer

Description automatically generated with low confidence**

Figure 34

Figure 33 shows the results of the k = 30, the accuracy is 0.6224, sensitivity is 0.10612 and the specificity is 0.91665. So, this model can predict borrower with pay-off very well.

**Model 2 (C.50 classification tree)**

A screenshot of a computer

Description automatically generated with low confidence

Figure 35

Figure 34 shows the results of C.50 classification tree, the accuracy is 0.7864 which is way better than k-nn model. The sensitivity and specificity are 0.6768 and 0.8488, both of them have a percentage higher than 50 which is more in line with the results I want

**Model 3 (random forest)**

A screenshot of a computer

Description automatically generated with low confidence

Figure 36

Figure 35 shows the result of random forest. Accuracy, sensitivity, and specificity are all slightly higher than the normal tree which make by C.50.

**Results**

|  |  |  |  |
| --- | --- | --- | --- |
|  | K-nn | C.50 classification tree | Random forest |
| **Accuracy** | **0.6224** | **0.7864** | **0.7941** |
| **Sensitivity** | **0.10612** | **0.6768** | **0.6782** |
| **Specificity** | **0.91665** | **0.8488** | **0.8601** |

From the results of these three models, random forest did the best, highest accuracy and sensitivity and with a specificity higher than 0.85. So, let’s try to use this model to predict the status on the origination time data.

Chart, bar chart

Description automatically generated

Figure 37

On the left side of figure 16, it is the prediction result, and the right side is the results in our original dataset. We can see there is a huge difference in predicting the borrower who default at the end.

Graphical user interface, text

Description automatically generated with medium confidence

Figure 38

Figure 37 shows the percentage of status time of borrowers in both results at the end. The first result is the prediction, and the second result is based on the dataset. The prediction put most of the borrower into pay-off, so this model may make the loan company to bankruptcy.

There is a possibility that we may need to treat the dataset as a time series by dividing it for each borrower, how much they pay for each period may give us more information for predicting if they will pay-off at the end.

**Case Study 3**

**Human Resource Descriptive Analytics**

Chun Ning Yip

Webster University

CSDA 6010

Dr. Ali Ovlia

November 9th, 2022

**Abstract**

There is always market competition between companies. Before twentieth century, companies need only the best machines and technology to dominate an industrial sector, but it is not enough for nowadays. Companies need also to considerate the importance of employees. Therefore, they would like to understand the reason why they are leaving or not satisfy with the company to maintain the productive employees.

In this project, the human resources analytics dataset has been used to do descriptive analytics and models building. We are going to explore the data and prove three hypotheses about the reason why employees want to leave the company first. Using techniques to find out the meaning between the data, then graphs will be plotted to show a clear sight between variables. After that we will build models and interpretate the results. The goal is to find out the workers if they will leave or stay in the company.

**Attribute interpretation**

This dataset contains ten attributes, satisfaction\_level and Last\_ evaluation is in numeric between 1 and 0. Number\_ project, average\_monthly\_hours and time\_spend\_company are also in numeric but with integers presenting the numbers. Work\_accident, left and promotion last 5years are categorical with 1 presenting yes and 0 presenting no. Lastly, sales and salary are categorical but in characters.

|  |  |
| --- | --- |
| **Satisfaction\_ level** | Level of satisfaction of employee |
| **Last\_ evaluation** | Last evaluation |
| **Number\_ project** | Number of projects |
| **Average\_ monthly\_ hours** | Average monthly hours |
| **Time\_ spend\_ company** | Time spend at the company |
| **Work\_ accident** | Whether they have had a work accident |
| **left** | Whether the employee has left |
| **Promotion\_ last\_ 5years** | Whether had a promotion in the last 5 years |
| **sales** | Departments (column sales) |
| **salary** | Salary |

A picture containing text

Description automatically generated

Figure 39

Table

Description automatically generated

Figure 40

Figure 39 shows the first five rows and last five rows of the dataset. In these 10 employees, most of them has a low satisfaction level and salaries, also they don’t appear in any work accidents. On our target variables “left”, it shows that all of them has left the company. So, it gives us a possible reason why employees left the company, low salaries, and satisfaction level.

**Data exploration**

In this dataset, there is no missing value inside any variables but the variable names and the type of variable need to be changed. In figure 40, left side shows the original column and the right-side shows columns after modified. On the left side, we can see sales department were under sales which doesn’t make sense, so I changed the column name into Department. Second, the salary was changed from categorical to numerical.

Shape, arrow

Description automatically generated

Figure 41

The target variable is in column “left”, the percentage of employees left vs stay is shown in figure 41. 1 means left and 0 means stay, there are 23.8% of employees left which is around one- third of the employees who stay.

Chart

Description automatically generated

Figure 42

Figure 42 shows the number of each department. Most of the employee is in Sales, support, and technical department. The rest of the apartment has only less than 1200 employees.

Chart, bar chart

Description automatically generated

Figure 43

Figure 43 shows the number of each salary types. Color yellow, green, and blue represents high, medium, and low respectively. Most of the employees falls into medium and low salaries.

Chart

Description automatically generated

Figure 44

Figure 44 shows the salaries in each department. Red is low salary type, green is medium salary type and blue is high salary type. Most of the employee have high salary were in management, sales, and technical department.

Chart, bar chart

Description automatically generated

Figure 45

Figure 45 shows box plots of three variables *satisfaction*, *last\_evaluation* and *number\_project*. Most of the employee left has low satisfaction, high evaluation, and high number of projects.

Chart, box and whisker chart

Description automatically generated

Figure 46

Figure 46 shows variables *average\_monthly\_hours* and *time\_spend\_company* by two box plots. The 1 and 0 on the x-axis represent left and stay respectively. Employee who left the company has a higher average monthly hour and has spent more time in the company.

Chart, box and whisker chart

Description automatically generated

Figure 47

Chart, bar chart, waterfall chart

Description automatically generated

Figure 48

Figure 47 shows four bar charts, from left to right, top to bottom represent work accident, promotion, department, and salary with the number of employees left and stay. From the graph, work accident and department are irrelevant, more employees left without accident and the ratio of employees left between each department are similar. Employee without promotion and with low or medium salaries has a higher chance on leaving the company.

**Hypothesis 1 (Salary is the reason why the employees left the company)**

Chart, bar chart

Description automatically generated

Figure 49

Figure 48 shows the percentage of left vs stay under each salary type. The percentages of employee left from low salary to high salary are 29.7%, 20.4% and 6.6% respectively. This means that salary is the reason why the employees left.

**Hypothesis 2 (Employees leave the company because work is not safe)**

Chart, bar chart

Description automatically generated

Figure 50

Figure 49 shows the percentage of employee left or stay by if they have work. The red bar shows the number of employees stay and blue bar shows the opposite. The percentage of employees who have work accident and left is lower than the employees who have no work accident and left. So, employees didn’t leave the company because work is not safe.

**Hypothesis 3 (Is this company a good place to grow professionally)**

There are 5 criteria need to be considerate. Satisfaction level, showing us how satisfy or good at their job. Average monthly hours divided by number of projects they have done, showing us how efficiency they are. Work accident, showing us if this company is safe, will there be high risk in having injuries. Promotion in last 5 years, showing us if they can get promote and be in touch with more work or area. Last evaluation, showing us how the company satisfy with the employee.

Chart, histogram

Description automatically generated

Figure

Most of the employees has a higher satisfaction level than 0.613 which is higher than the average, so the first criteria pass, they are satisfy and good in the job.

Chart, histogram

Description automatically generated

Figure

Figure 51 is a density plot of the average monthly hours divided by number of projects they have done. There are two peaks around the mean, one is slightly higher, and one is slightly lower, which means that most of the employees are so far so good. So, the second criteria will not be count.

Chart

Description automatically generated

Figure

Figure 52 shows the number of employees have or not have work accident. More than 80% of employee don’t have work accident, so most of them are safe, so criteria 3 pass.

Chart, bar chart

Description automatically generated

Figure 54

Figure 53 shows the number of employees get promotion in last 5 years. More than 97.9% of employee has no promotion in the last 5 years. Which is really low, so criteria 4 fail.

Chart, histogram

Description automatically generated

Figure 55

Figure 54 is a density plot showing the last evaluation. The area over the mean is larger than below the mean, which tell us the company are highly satisfied with more than half of employees. So, criteria 5 pass.

In conclusion, this company pass four criteria and failed one, so I would say it is fine to grow professionally in this company but maybe after you have some experience employee should hop into another company to grow professionally further, since the promotion rate in last 5 years is extremely low.

**Data Partitioning**

Starting with selecting important variables by using logistic regression. Figure 55 is the result of logistic regression, some of the departments are not important but we can still keep it. Then the dataset is separated 60% to training and 40% to validation. C.50 and rpart will be used to build the classification tree.

A screenshot of a computer

Description automatically generated with low confidence

Figure

**Models’ Results**

|  |  |
| --- | --- |
| C.50 | rpart |
|  |  |

In both results, their accuracy, sensitivity and specificity are very high and similar, all of them are above 90%. This model maybe overfitting to the training set, if there is a testing set, I will know the answer and tune the model.