CS 6375.001 Machine learning

Name: Bhavya Thadiboina

NET_ID: BXT220009

Question 1:

Determining dependent and independent variables

The outcome or endogenous variable is Satisfaction level and the input variables are the rest of the variables excluding Employee ID. If we are to write it as a linear function, we will have:

$S=\beta Tj\ \dot{}Xi+\beta 0$

Where,
S = Satisfaction Level
Xi = independent variables
βj = Weights of the various independent variables.
$\beta 0$ = Bias

Summary statistic of numerical columns

The satisfaction level statistic

The satisfaction level statistic provides an understanding of the overall satisfaction of the employees in the company. The average satisfaction level of 0.613 suggests a moderate level of satisfaction, with some variations around this average indicated by a standard deviation of 0.249. The quartiles reveal further information: 25% of the employees have a satisfaction level lower than 0.44, while 75% have a satisfaction level lower than 0.82. The median satisfaction level is 0.64, indicating that half of the employees have a higher satisfaction level and half have a lower satisfaction level. The minimum and maximum values show the range of satisfaction levels observed, with the lowest recorded level being 0.09 and the highest being 1.0. The variance of 0.062 indicates the extent of dispersion or spread in the satisfaction levels.

The last evaluation statistic

The last evaluation statistic helps gauge the performance evaluation scores of employees. With an average score of 0.716 and a standard deviation of 0.171, the evaluations tend to be relatively positive overall. The quartiles provide additional insights: 25% of the employees have an evaluation score lower than 0.56, while 75% have an evaluation score lower than 0.87. The median evaluation score is 0.72. The minimum and maximum values of 0.36 and 1.0, respectively, demonstrate the range of evaluation scores. The variance of 0.029 indicates the variability in the evaluation scores.

The number of projects statistic

The number of projects statistics reveal the workload distribution among employees. The average number of projects assigned per employee is 3.803, with a standard deviation of 1.233, suggesting some variation in the workload. The quartiles provide further insights: 25% of the employees have fewer than 3 projects assigned, while 75% have fewer than 5 projects assigned. The median number of projects is 4. The minimum and maximum values of 2 and 7, respectively, represent the range of projects assigned. The variance of 1.519 reflects the spread in the number of projects.

The average monthly hours statistic

The average monthly hours statistic indicates the average time employees spend working each month. With an average of 201.050 hours and a standard deviation of 49.943, there is some variation in the monthly working hours. The quartiles offer additional information: 25% of the employees work fewer than 156 hours per month on average, while 75% work fewer than 245 hours. The median monthly working hours is 200. The minimum and maximum values of 96 and 310, respectively, represent the range of hours worked. The variance of 2494.313 demonstrates the dispersion in the monthly working hours.

The time spent in the company statistic

The time spent in the company statistic reflects the tenure or duration of employees' employment. The average tenure is approximately 3.498 years, with a standard deviation of 1.460, indicating some variability in the duration of employment. The quartiles provide further insights: 25% of the employees have spent less than 3 years in the company, while 75% have spent less than 4 years. The median tenure is 3 years. The minimum and maximum values of 2 and 10 years, respectively, represent the range of tenures. The variance of 2.132 indicates the spread in the tenure durations.

The work accident statistic

The work accident statistic highlights the occurrence of accidents in the workplace. On average, there is a 0.145 probability of experiencing a work accident, with a standard deviation of 0.352, indicating some variability in the occurrence of accidents. The quartiles provide additional information: 25% of the employees have not experienced a work accident, while 75% have not experienced a work accident. The median work accident count is 0, indicating that half of the employees have not experienced a work accident. The minimum and maximum values of 0 and 1, respectively, represent the range of work accident counts. The variance of 0.124 shows the extent of dispersion in the occurrence of work accidents.

The promotion in the last 5 years statistic

The promotion in the last 5 years statistic provides insights into the likelihood of employees receiving promotions. On average, there is a 0.021 probability of receiving a promotion in the last 5 years, with a standard deviation of 0.144, indicating limited occurrences of promotions. The quartiles offer further insights: 25% of the employees have not received a promotion in the last 5 years, while 75% have not received a promotion in the last 5 years. The median promotion count is 0, indicating that half of the employees have not received a promotion in the last 5 years. The minimum and maximum values of 0 and 1, respectively, represent the range of promotion counts. The variance of 0.021 reflects the spread in the occurrence of promotions.

Summary statistic of categorical columns

The dataset contains information for **14,999** employees, with **10** unique departments and **3** unique salary categories; the most common department is sales with a frequency of **4,140** and the most common salary category is low with a frequency of **7,316.**

Question 2:

Detecting Outliers with Inter Quartile Range (IQR)

Interpretation of the results

In boxplots, any number greater the upper limit (upper limit=Q1−1.5∗(Q3−Q1)) 1−1.5∗(3−1)) or less than the lower limit (lower limit=Q3+1.5∗(Q3−Q1))=3+1.5∗(3−1)) is considered and Outliers

From the boxplots above, all the data values with asterisk (*) are outliers.

Question 3:

Analysis of the Counts of categorical columns

Interpretation of the results

From the bar chart on the left above, sales department has the high number of employees **(4140)**, followed by technical department **(2720)** etc.

From the bar chart on the right above, most employees receive low pay **(7316)**, and the number of employees that receive high salary is very low, **(1237)**.

Question 4:

Interpretation of the results

1. **Satisfaction Level:**

Employees with higher satisfaction levels tend to have higher salaries. The average satisfaction level is highest for high salaries (0.637), followed by medium salaries (0.622), and lowest for low salaries (0.601).

2.Last Evaluation:

While not as strong as satisfaction level, the last evaluation score shows some influence on salary level. However, the average scores for last evaluation are similar across low, medium, and high salaries.

2. Number of Projects:

The number of projects assigned to employees does not seem to strongly predict salary level. The average number of projects is relatively consistent across all salary levels.

3. Average Monthly Hours:

The average monthly hours worked by employees do not appear to be a strong predictor of salary level. There is no clear trend indicating that longer hours result in higher salaries.

4. Time Spent in the Company:

The length of time spent in the company does not show a significant association with salary level. The average time spent in the company is comparable across low, medium, and high salary levels.

5. Work Accident and Promotion:

Work accidents and promotions in the last five years show slight variations across salary levels, but their impact on salary determination is not prominent based on the provided data.

In conclusion, based on the observed associations, the satisfaction level appears to be the strongest predictor of salary level, followed by last evaluation. The number of projects, average monthly hours, and time spent in the company have less influence on salary determination. Work accidents and promotions show limited influence and require further investigation.

Question 5:

The relationship between the the variables can be estimated using the pearson correlation coefficient. he correlation coefficient measures the strength and direction of the linear relationship between two variables.

$\rho(x,y)=\sum(xi-x^{-})(yi-y^{-})/\sqrt{\sum(\mathbf{xi}-\mathbf{x}^{-})2\sum(yi-y^{-})\mathbf{2}}$

Interpretation of the results

The correlation coefficient measures the strength and direction of the linear relationship between two variables. In this case, we are examining the correlation between several variables and the satisfaction level. Let's interpret the correlation coefficients one by one:

- **last_evaluation (0.105):** There is a weak positive correlation between last_evaluation and satisfaction level. This suggests that employees who have higher last evaluation scores may tend to have slightly higher satisfaction levels, although the relationship is not very strong.
- **Work_accident (0.059):** There is a very weak positive correlation between work accidents and satisfaction level. This means that employees who have experienced work accidents might have slightly higher satisfaction levels, although the correlation is almost negligible.
- **promotion_last_5years (0.026):** There is a very weak positive correlation between promotions in the last five years and satisfaction level. This suggests that employees who have received promotions in the past five years might have slightly higher satisfaction levels, although the correlation is minimal.
- **average_montly_hours (-0.020):** There is a very weak negative correlation between average monthly hours and satisfaction level. This implies that employees who work longer hours may have slightly lower satisfaction levels, although the correlation is almost negligible.
- **time_spend_company (-0.101):** There is a weak negative correlation between the duration of employment (time spent at the company) and satisfaction level. This means that employees who have been with the company for a longer time might have slightly lower satisfaction levels, although the relationship is not very strong.
- **number_project (-0.143):** There is a moderate negative correlation between the number of projects an employee has worked on and satisfaction level. This indicates that employees who are assigned to a higher number of projects may tend to have lower satisfaction levels, suggesting a somewhat stronger relationship compared to the other variables.

Question 6:

Interpretation of the results

- **Highest Average Satisfaction Level**

The highest average satisfaction level among employees is observed in the HR department, regardless of salary level, with a satisfaction score of 0.673 for high salary, 0.609 for low salary, and 0.580 for medium salary.

- **IT and Sales Departments**

  In general, employees in the IT and sales departments have relatively high average satisfaction levels across all salary levels, with scores ranging from 0.638 to 0.655 for high salary, 0.594 to 0.656 for low salary, and 0.620 to 0.645 for medium salary.

- **Management and Marketing Departments**

  The management and marketing departments also show a relatively positive average satisfaction level, with scores ranging from 0.605 to 0.653 for high salary, 0.603 to 0.619 for low salary, and 0.597 to 0.638 for medium salary.

- **Accounting and Product Management Department**

  Employees in the accounting and product management departments tend to have slightly lower average satisfaction levels compared to other departments, with scores ranging from 0.574 to 0.614 for low salary, 0.584 to 0.620 for medium salary, and 0.586 to 0.615 for high salary.

- **Support and Technical Departments**

  The support and technical departments exhibit moderate average satisfaction levels across all salary levels, with scores ranging from 0.592 to 0.655 for low salary, 0.620 to 0.645 for medium salary, and 0.626 to 0.649 for high salary.

Interpretation of the results

From the plot above there is no clear linear relationship among the variables. Also, all the varibles have deviated from normal distribution; almost all of them are showing multimodal distribution.

Question 7:

In this section, we build a Decision Tree Regressor to predict an employee satisfaction level based the independent variables in the dataset. Also, the model is used to determine the best predictors of employee satisfaction

The Mean Square Error (MSE=1\n ∑ni=1(Yi−Y^i)^2 is used as the metric for determining the best split and the overall performance of the model.

Interpretation of the results

The coefficient of determination is **39.51%**, indicating **39.51% of the variation in satisfaction level is explained by variations in the independent variables**