



Project Name

EDA and Statistical Analysis of Credit Card Dataset

—

Overview

This Statistics and EDA project is designed to train and test you on basic Data Exploratory and Statistical techniques used in the industry today. Apart from bringing you to speed with basic descriptive and inferential methods, you will also deep dive into a dataset and perform thorough cleaning and analysis in order to draw useful business insights from the data. This will expose you to what data scientists do most often—Exploratory Data Analysis.

Goals

1. Using the core statistical theoretical concepts and knowledge to solve real time problem statements.
2. Visualize a real time industry scenario where one can use these statistical concepts.
3. Detailed data analysis and number crunching using statistics
4. Exhaustive report building using EDA and visualization techniques to help the business take decisions using insights from the data

Specifications

Part -I is concept based and walks you through various concepts of descriptive statistics, probability distributions and inferential statistics including confidence intervals and hypothesis testing.

Part -II on the other hand is dataset based and explore various data cleaning options, data analysis options and using EDA to derive deep and meaningful insights for the business

PART-A (Concept Based)--25 points

The following are the ages of CEOs of 42 Top Fortune 500 Companies when they took over the post of CEO

57 61 57 57 58 57 61 54 68

51 49 64 50 48 65 52 56 46

54 49 50 47 55 55 54 42 51

56 55 54 51 60 62 43 55 56

61 52 69 64 46 54

Use this data for answering following questions where relevant.

Q1. Compute the mean, median and the mode of the data

Q2. Compute the range , variance and standard deviation of CEO ages

Q3. Find the mean deviation for the data . The mean deviation is defined as below.

$$\text{Mean deviation} = \frac{\sum |X - \bar{X}|}{n}$$

Q4. Calculate the Pearson coefficient of skewness and comment on the skewness of the data

[A measure to determine the skewness of a distribution is called the Pearson coefficient of skewness. The formula is

$$\text{Skewness} = \frac{3(\bar{X} - MD)}{s}$$

where MD is the median and s the standard deviation

The value of the coefficient of skewness usually ranges from –3 to 3. When the distribution is symmetric, the coefficient is zero; when the distribution is positively skewed , the coefficient is positive, and when the distribution is negatively skewed the coefficient is negative.]

Q5. Count the number of data values that fall within two standard deviations of the mean. Compare this with the answer from Chebyshev's Theorem.

Q6. Find the three quartiles and the interquartile range (IQR).

Q7. Are there any outliers in the data set ? Q8.

Draw a boxplot of the dataset to confirm .

Q9. Find the percentile rank of the datapoint 50.

Q10. What is the probability that a person becoming a CEO is below 50 years of age ?

Q11. Create a frequency distribution for the data and visualize it appropriately Q12.

Create a probability distribution of the data and visualize it appropriately. Q13. What

is the shape of the distribution of this dataset? Create an appropriate graph to

determine that. Take 100 random samples with replacement from this dataset of

size 5 each. Create a sampling distribution of the mean age of customers. Compare

with other sampling distributions of sample size 10, 15, 20, 25, 30. State your

observations. Does it corroborate the Central Limit Theorem?

Q14. Treat this dataset as a binomial distribution where p is the probability that a

person becomes a CEO above 50 years age. What is the probability that out of a

random sample of 10 CEOs of Fortune 500 companies exactly 6 are above 50 years

of age?

Q15. A study claims that 25% of all Fortune 500 companies becoming a CEO are above

60 years of age. Using the Normal approximation of a Binomial distribution, find the

probability that in a random sample of 300 Fortune 500 companies exactly 75 CEOs

will be above 50 years of age.

[Note that the normal distribution can be used to approximate a binomial distribution if $np \geq 5$ and $nq \geq 5$ with the following correction for continuity

$P(X=z) = P(z-0.5 < X < z+0.5)]$

Q16. Compute a 95% Confidence Interval for the true mean age of the population of

CEOs for the given dataset using appropriate distribution. (State reasons as to why

did you use a z or t distribution)

Q17. A data scientist wants to estimate with 95% confidence the proportion of CEOs

of Fortune 500 companies are above 60 years in the population.

Another recent study showed that 25% of CEOs interviewed were above 60. The

data scientist wants to be accurate within 2% of the true proportion. Find the

minimum sample size necessary.

Q18. The same data scientist wants to estimate the true proportion of CEOs

ascending to the post and above 60 years. She wants to be 90% confident

and accurate within 5% of true proportion. Find the minimum sample size necessary.

Q19. A researcher claims that currently 25% of all CEOs are above 60 years .Test his claim with an $\alpha = 0.05$ if out of a random sample of 30 CEOs only 10 are above 60 years.

Q20. Assume you are a data scientist for the Fortune 500 companies. You are asked to research the question whether the CEO ages of UK are on average older than the CEO ages of Americans. you take a random sample of 40 CEO ages from America and UK and the data is as follows:

UK

47 49 73 50 65 70 49 47 40 43

46 35 38 40 47 39 49 37 37 36

40 37 31 48 48 45 52 38 38 36

44 40 48 45 45 36 39 44 52 47

USA

47 57 52 47 48 56 56 52 50 40

46 43 44 51 36 42 49 49 40 43

39 39 22 41 45 46 39 32 36 32

32 32 37 33 44 49 44 44 49 32

- What are your hypotheses?
- What significance level will you use?
- What statistical test will you use?
- What are the test results? (Assume $s_1 = 8.8$ and $s_2 = 7.8$.)
- What is your decision?
- What can you conclude?
- Do you feel that using the data given really answers the original question asked?
- What other data might be used to answer the question?

PART-B (Dataset Based)--25 points

Topic - Credit Card Fraud Detection

Introduction

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

Dataset Description


The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions.

Data Dictionary

- a) It contains only numeric input variables. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'.
- b) Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset.
- c) The feature 'Amount' is the transaction Amount,
- d) Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Questions -

1. Import the dataset and view the first 10 rows of it.
2. Display shape/dimension of the dataset.
3. Check for the missing values. Display number of missing values per column.
4. Check the datatype, number of non-null values and name of each variable in the dataset.
5. Check if there are any non-real characters in the dataset.
6. Check the descriptive statistics of the dataset.

- 
7. Check the number of fraudulent transactions in the dataset and visualize using pie chart and bar plot.
 8. Check the maximum and minimum fraudulent amount.
 9. Check the number of transactions where the transaction amount is zero and consider as a fraud transaction.
 10. Check the distribution of columns. List down columns that are normally distributed. List down columns that are not normally distributed.
 11. List down columns that are highly skewed.
 12. With the help of a standard scaler, normalize the respective column distribution.
 13. List down columns that have high kurtosis.
 14. What is the distribution of Time and Amount columns in the dataset ?
 15. Find the distribution of all variables with respect to the outcome 'Class' variable.
 16. Create a countplot for the outcome class in seaborn using percentage instead of count for each bar.
 17. Plot a heatmap for correlation matrix for the given dataset. Write the Observation. Especially note down columns that are highly correlated (Positive and Negative Correlation, Consider 0.7 to 1 as high).
 18. With the help of hypothesis testing check whether fraudulent transactions of higher value than normal transactions?
 19. Perform ANOVA test for Statistical feature selection.
 20. Split the dataset randomly into train and test datasets. Use a train -test ratio of 70:30 ratio.
 21. These are just checkpoints. Please use your best analytical approach to build this report. You can mix match columns to create new ones which can be

used for better analysis. Create your own features if required. Be highly experimental and analytical here to find hidden patterns. You can use the following as checklist pointers :

- What is the shape and size of the dataset?
 - Which columns are highly skewed?
 - Which columns are highly Kurtosis driven?
 - Which columns have Wrong data type?
 - What columns seem to have outliers based on min, max and percentile values, IQR range along with the standard deviation and mean absolute deviation?
 - What columns have missing values? (Check the Missing Values section in Pandas Profiling)
 - What columns have high amount of zero and make sure that these zeroes are supposed to be there(for eg. Weight cannot be zero and any percentage of zero in column zero is erroneous)
 - What columns have high variance and standard deviation?
 - Comment on the distribution of the continuous values (Real Number: $\mathbb{R} \geq 0$)
 - Do you see any alarming trends in the extreme values (minimum 5 and maximum 5)?
 - How many Boolean columns are there in the data set and out of those how many are imbalanced?
 - Check for duplicate records across all columns (Check Warning Section)
 - Is there any imbalance in the categorical columns? (for example Gender Male and Female in which Male is 95% and Female is just 5%- How many columns are categorical?)
 - Are those categories in sync with the domain categories?
 - Check if all the categories are unique and they represent distinct information
- Based on the above questions and your observations, chart out a plan for Data Pre-processing and feature engineering

Note: Feature Engineering (Feature Selection and Feature Creation)