Data Analysis Portfolio

Prepared By:-
Bishwomkar Panigrahi

Professional Background

Currently in my third year pursuing B.Tech-CSE. I have secured 8.48 CGPA (till 5th sem) and have several skills including Data Analysis, JAVA, C, C++.

I have worked with several companies as an intern like Eventbeep, Internshala etc, as a student ambassador as well as marketing strategist. I have also worked on my social skills as well as communication skills for better people management to lead a team, moreover I have done several course from Linkedin learning on time management to get an efficient balance.

I was also part of my college IEEE group where I learnt a great amount of new skills and also worked closely with many hardware projects based on Iot. Ihave recently completed my R programming course from Infoysys learning site.

As I am a fresher it would be great to experience the real challenges of the corporate world and understand how things work. Being a fresher, I think I am very flexible and adaptive to learn new things. I have theoretical knowledge. But I am waiting to use my theoretical knowledge in a practical way. And I believe by putting significant efforts I will learn.

Table Of Contents

Data Analytics Process

Description: We use Data Analytics in everyday life.
For example: Searching in Web Searching Engines like Google.

6 Steps Process:

Plan:
We first decide which things we need to search before opening the Google. Which information do we need to search? For example, we want to know about different machine learning algorithms.

Prepare:
Next we need to check which website will give us the correct information in the simplest way.

Process:
Then we need to check how much we want from the data. Like if we want – difference between the different algorithms / which algorithm is the best choice according to the different scenarios / etc.

Analyze:
We then analyze the different algorithms. Suppose we are working on some project and we don't know how to apply the algorithms or which algorithm will be the best choice in that project.

Share:
Now we search it in the Google. And Google gives us the best results at the top-most by using the process of data analytics / data science.

Act:
Then we finally click it and get the necessary information which we need. We, then use that information in our project.

Project Link:

https://drive.google.com/file/d/1aYv2oRgOMsvBJhzs8VzNiJj6PA0Dqrgh/view?usp=drivesdk

Instagram User Analytics

Description: This project is about how the users engage and interact with Instagram. We will analyze these users in an attempt to derive business insights for marketing, product & development teams. These insights are then used by teams across the business to launch a new marketing campaign, decide on features to build for an app, track the success of the app by measuring user engagement and improve the experience altogether while helping the business grow.

Findings:

- **Rewarding Most Loyal Users:** The 5 oldest users of the Instagram
- **Remind Inactive Users to Start Posting:** The users who have never posted a single photo on Instagram
- **Declaring Contest Winner:** The user who gets the most likes on a single Photo.
- **Hashtag Researching:** The top 5 most commonly used hashtags
- **Launch AD Campaign:** Day of the week do most users register on
- **User Engagement:** Average user posts on Instagram
- **Bots & Fake Accounts:** Users (bots) who have liked every single photo on the site.

Approach: We are working with the product team of Instagram and the product manager has asked us to provide insights on the questions asked by the management team. We use SQL to derive different insights from the dataset provided by the management team. First, we run the necessary commands for creating the database to work on. Then, we performed analysis to generate valuable insights for the company.

Insights:

There are total of 100 users using Instagram clone.

Around 26% of the users are inactive in Instagram. We can remind the inactive users by sending them promotional emails to post their 1st photo.

The most liked photo in Instagram is posted by Zack_Kemmer93, which is liked by 48% of the users. The team can start the contest for the most liked photos. This will make the users to post more such good posts.

The most used hashtag is "smile". Around 59% of the users use the "smile" hashtags. If a partner brand use the "smile" hashtag, it will be able to reach the most users in the platform.

The best days to launch ads are Sunday and Thursday. As the most users register on Instagram on Sunday and Thursday.

13% of Instagram IDs are fake and dummy accounts.

ProjectLink:

https://drive.google.com/file/d/1Tt5_5ILyGjNxRuwEeRsN1SkJQ0N2Nhfl/view?usp=drivesdk

Operation Analytics & Investigating Metrics

Description: Operation Analytics is the analysis done for the complete end to end operations of a company. With the help of this, the company then finds the areas on which it must improve upon. Being one of the most important parts of a company, this kind of analysis is further used to understanding between cross-functional teams, and more effective workflows.

Investigating metric spike is also an important part of operation analytics as being a Data Analyst we must be able to understand or make other teams understand questions like- Why is there a dip in daily engagement? Why have sales taken a dip? Etc. Questions like these must be answered daily and for that it's very important to investigate metric.

Findings:

- **Number of jobs reviewed**
- **Throughput**
- **Percentage share of each language**
- **Duplicate rows**
- **User Engagement**
- **User Growth**
- **Weekly Retention**
- **Weekly Engagement**
- **Email Engagement**

Approach: I am working for a company like Microsoft designated as Data Analyst Lead and is provided with different data sets, tables from which I must derive certain insights out of it and answer the questions asked by different departments. Firstly, I spent some time on understanding the data/table given. I cleared the questions which was in my mind and what are the things to consider while reviewing the data. I use SQL to derive different insights from the dataset provided by the management team. I first created a database "operation_analytics" and then the tables using the structure and links provided by the team. Then, we performed analysis to generate valuable insights for the company.

Insights:

1. Case Study 1 (Job Data):

- The number of distinct jobs reviewed per hour per day for November 2020 is 83%.
- We used the 7-day rolling average of throughput as it gives the average for all the days right from day 1 to day 7 whereas, daily metric gives the average for only that particular day itself.
- The percentage share of Persian language is the most (37.5%).
- There are two duplicate rows if we partition the data by job_id. But if we look the overall columns, all the rows are unique

2. Case Study 2 (Investigating metric spike):

- The weekly user engagement increased from week 18th to week 31st and then started declining from then onward. This means that some of the users do not find much quality in the product/service in the last of the weeks.
- There are in total 9381 active users from 1st week of 2013 to the 35th week of 2014.
- The overall count of weekly engagement per device used is the most for MacBook users and iPhone users.
- The email opening rate is around 34% and email clicking rate is around 15%. The users are engaging with the email service which is good for the company to expand

Project Link:

https://drive.google.com/file/d/1PrcZmYHBo5II2V93uCJGtr83dieN3ny/view?usp=drivesdk

Hiring Process Analytics

Description: Hiring process is the fundamental and the most important function of a company. Here, the MNCs get to know about the major underlying trends about the hiring process. Trends such as- number of rejections, number of interviews, types of jobs, vacancies etc. are important for a company to analyze before hiring freshers or any other individual. Being a Data Analyst, our job is to go through these trends and draw insights out of it for hiring department to work upon.

Findings:

- **Hiring:** How many males and females are Hired ?
- **Average Salary:** What is the average salary offered in this company ?
- **Class Intervals:** Draw the class intervals for salary in the company ?
- **Charts and Plots:** Draw Pie Chart / Bar Graph (or any other graph) to show proportion of people working different department ?
- **Charts:** Represent different post tiers using chart/graph?

Approach: I am working for a MNC such as Google as a lead Data Analyst and the company has provided with the data records of their previous hirings and have asked me to answer certain questions making sense out of that data. We will use EDA to generate different insights and to answer the questions asked by the company. The dataset given by the company contains the details about people who registered for a particular post in a department of this company. I used MS Excel to analyze the data with different tables and columns.

Insights: The rejection rate of male applicant is 6% higher than the female applicant.
The average salary paid in this company is 50K.
Most of the employers are in the Operation Department and then in the Human Resource Department.
The applicant is most likely to get hired if he/she is applying for the HR Department as the rejection rate here is the least.
There are only 3 candidates in the company who are paid more than 100K.

Project Link:

https://drive.google.com/file/d/1NdkuBiCQQDOmask69cJZeUHICNRuM8b/view?usp=drivesdk

IMDB Movie Analysis

Description: The dataset provided by the company contains various columns of different IMDB Movies. We are required to Frame the problem. For this task, we will need to define a problem we want to shed some light on.
We can do this by asking the following 'What?' :

- What do we see happening?
- What is our hypothesis for the cause of the problem? (this will be broadly
- based on intuition initially)
- What is the impact of the problem on stakeholders?
- What is the impact of the problem not being solved?

Findings: The things that we find out through the project are:

- movies with the highest profit
- top movies as per imdb rating
- top directors
- most popular genres
- top foreign language films

Approach: Firstly, I cleaned the data. Then, we used Five 'Whys' approach to determine its root cause by repeatedly asking the question "Why". While asking Why is easy, what we're interested in is the answer. Each time we answer why the next time gets more difficult as we must think deeper behind the reasons for this. As we ask why, we may find that we have multiple answers for the same question.

Insights:

- There are as many as 5 outliers in the profit columns.
- The movie with the highest profit is 'Avatar' followed by 'Jurassic World' and 'Titanic' and so on.
- The Shawshank Redemption is the top-most movie with the highest IMDB rating.
- The Good, the Bad and the Ugly (Italian) is the top-most foreign language movie.
- Charles Chaplin is the top-most director followed by Tony Kaye.
- The most popular genres is Drama followed by Comedy.
- 'Leonardo DiCaprio' is the critic-favorite as well as the audience-favorite actor.
- The most users voted in the decade 2000s and the least in the decade 1940s.

Project Link:

https://drive.google.com/file/d/1OB5nYpfIP3lafDZg9PTQaiv2g3GAyYmk/view?usp=drivesdk

Bank Loan Case Study

Description: This case study aims to give us an idea of applying EDA in a real business scenario. In this case study, we will develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Findings:

- Our aim is to identify the patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- The driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.
- Presenting the overall approach of the data analysis, cleaning the dataset, finding outliers, data imbalance, univariate, segmented univariate, bivariate analysis, etc.
- The top 10 correlation for the Client with payment difficulties and all other cases (Target variable).

Approach:

- Imported the NumPy, pandas, matplotlib and seaborn python libraries.
- Imported the datasets (Application_Data & Previous_Application)
- Identification: We have identified how we will approach the data, finding missing dataset and working on it accordingly to gain the required results.
- Outliers: Identified outliers and showed how they play if any role in our dataset.
- Imbalance: Understanding the ratio of imbalance in our data.
- Correlation Analysis: Finding the correlation between the variables with respect to the target variables and find the top three correlation.
- Visualization: Visualized the data with the help of charts and graphs.

Insights:

1. Non-Default:

- Academic degree has less defaults.
- Student and Businessmen have no defaults.
- Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%.
- People above age of 50 have low probability of defaulting.
- Applicant with Income more than 700,000 are less likely to default.
- People with zero to two children tend to repay the loans.

2. Default:

- Men are at relatively higher default rate.
- Clients who are either at Maternity leave OR Unemployed default a lot.
- Not approving the loan of young people who are in age group of 20-40 as they have higher probability of defaulting.
- When the credit amount goes beyond 3M, there is an increase in defaulters.
- People who have less than 5 years of employment have high default rate.

Project Link:

https://drive.google.com/file/d/10AQIsUuN_6YIrpGQ6bOoR8yZsPAzl25F/view?usp=drivesdk

XYZ Ads Airing Report Analysis

Description: Advertising is a way of marketing our business in order to increase sales or make our audience aware of our products or services. Until a customer deals with us directly and actually buys our products or services, our advertising may help to form their first impressions of our business. Target audience for businesses could be local, regional, national or international or a mixture. So they use different ways for advertisement. Some of the types of advertisement are: Internet/online directories, Trade and technical press, Radio, Cinema, Outdoor advertising, National papers, magazines and TV. Advertising business is very competitive as a lot of players bid a lot of money in a single segment of business to target the same audience. Here comes the analytical skills of the company to target those audiences from those types of media platforms where they convert them to their customers at a low cost.

Findings:

- What is Pod Position? Does the Pod position number affect the amount spent on Ads for a specific period of time by a company?
- What is the share of various brands in TV airings and how has it changed from Q1 to Q4 in 2021?
- Conduct a competitive analysis for the brands and define advertisement strategy of different brands and how it differs across the brands.
- Mahindra and Mahindra wants to run a digital ad campaign to complement its existing TV ads in Q1 of 2022. Based on the data from 2021, suggest a media plan to the CMO of Mahindra and Mahindra. Which audience should they target?

Approach:

- Scatter chart w.r.t different brands is used to know if the pod position affects the amount spent on Ads for a specific period of time by the company.
- Bar chart and column chart are used for answering the share of various brands in TV Airings.
- We used pivot table to conduct the competitive analysis for the brands.
- Clustered Column chart is used to suggest a media plan to the CMO of Mahindra and Mahindra

Insights:

- The brand's money spent for the advertisement is the least for the last quarter pod position and the highest for the first quarter pod position.
- The money spent by Mahindra and Mahindra is the most for the pod position ads.
- The money spent by Honda Cars is the least for the pod position ads.
- The money spent by the Maruti Suzuki is the most consistent for all the Quarters of the year.
- People watch the most in the prime time and on weekend.
- The Ads are shown the least in the prime access and evening news parts of the day

Project Link:

https://drive.google.com/file/d/1YvacLPc23bSJgtE4UmNrae_HGBKyrsKQ/view?usp=drivesdk

ABC Call Volume Trend Analysis

Description: The attached dataset is of Inbound calls of an ABC company from the insurance category consists of a Customer Experience (CX) Inbound calling team for 23 days. Data includes Agent_Name, Agent_ID, Queue_Time [duration for which customer have to wait before they get connected to an agent], Time [time at which call was made by customer in a day], Time_Bucket [for easiness we have also provided you with the time bucket], Duration [duration for which a customer and executives are on call, Call_Seconds [for simplicity we have also converted those time into seconds], call status (Abandon, answered, transferred)

Findings:

- The average call time duration for all incoming calls received by agents (in each Time_Bucket).
- The total volume/ number of calls coming in via charts/ graphs [Number of calls v/s Time].
- Propose a manpower plan required during each time bucket [between 9am to 9pm] to reduce the abandon rate to 10%.
- Propose a manpower plan required during each time bucket in a day[9 pm to 9 am]. Maximum Abandon rate assumption would be same 10%.

Approach:

- We used pivot table and pivot charts to get the valuable insights of the data.
- We assumed an agent work for 6 days a week;
- On an average total unplanned leaves per agent is 4 days a month; An agent total working hrs is 9 Hrs out of which 1.5 Hrs goes into lunch and snacks in the office.
- On average an agent occupied for 60% of his total actual working Hrs (i.e. 60% of 7.5 Hrs) on call with customers/ users.
- We also assumed total days in a month is 28 days for easy calculation

Insights:

- The customers call the least in the evening. So, the company can reduce the
- number of agents at that time for answering the calls.
- The company can hire 17 customer support agents for the night shift work.
- The company can shift some of the day workers for the night shift.
- The employees who are working 9 am to 9 pm. The manager can change some of the workers shift from 5 am to 2 pm and some workers from 2 pm to 11 pm to get the most calls answered.
- The company can make the employers divide into 3 parts too, so that the agents are always available 24/7.

Project Link:

https://drive.google.com/file/d/1Qzu_Hx7p1xpNahnF77J5GMaVbn6FLdMy/view?usp=drivesdk

CONCLUSION

Learnings: The things I learned from the projects are:

- Data Analysis 6 Steps Processes
- How to use Advance SQL Concepts in the real world business case
- How to use Advance Excel Concepts in the real business case scenario
- How to analyze the huge datasets in Python, Excel, etc.
- How to visualize the data to gain the valuable insights
- Concepts of Operation Analysis & Investigating Metric Spikes
- HR Analytics
- Predictive Analytics
- Risk Analytics
- Behavioral Analytics
- Business scenario of Ads Airing Report
- Customer Experience Team and Inbound Customer Support
- Google to get the concepts and answers whenever get stuck

Submitted by :- **Bishwomkar Panigrahi**