SUMMER INTERNSHIP COURSE

Group_Name: Grid Crafters-

1. Bidwai Gajanan Arun 12313438
2. Gampala Vijay Kumar 12304768
3. Musunuri Yajna Sri 12308182
4. Sashil Kumar Singh 12313595
5. Prudhvi 12306236

Submitted To – Sandeep Kaur Ma'am,

Jaffar Amin Chacket Sir

Title Of Project:

Employee Performance Segmentation for HR Insights

DECLARATION

I, Bidwai Gajanan Arun, a student of Lovely Professional University, pursuing B.Tech in Computer Science and Engineering/Information Technology, hereby declare that the work presented in this project report is the result of our group's sincere efforts and dedicationnder the team name GRID CRAFTERS. The information and findings documented herein are based on genuine research, analysis, and practical implementation carried out by us. This report is submitted as a part of the academic curriculum and is a true representation of our original work.

Date: 12/07/2025
Registration No. 12313438
Name of the student: Bidwai Gajanan Arun

1.Problem Statement:

Create a project that enables performance segmentation of employees based on productivity indicators such as task completion efficiency, attendance regularity, feedback ratings, and length of service. Employ unsupervised learning methods like clustering to identify distinct employee performance profiles, ranging from top performers to those needing support. Use Power BI to deliver insightful visualization that break down performance metrics by team, department, and tenure,aiding HR in making data-informed development plans

2. Project Objectives

This project aims to deliver a comprehensive data-driven system for HR analytics, with the following objectives:

- Identify performance patterns and segment employees using data analytics.
- Predict employee churn using machine learning techniques.
- Visualize performance metrics using Power BI dashboards.
- Utilize SQL queries to extract meaningful business insights.
- Assist HR professionals in making strategic decisions based on employee behavior

3. Dataset Description

The dataset used in this project is named 'Customer_Data.csv', which has been treated analogously to employee performance data. It contains key attributes such as:

- Customer_ID (Employee ID)
- Tenure_in_Months
- Monthly_Charge, Total_Charges
- Customer_Status (Active, Churned)
- Gender, Plan Type, Service Usage Metrics

The dataset is cleaned, preprocessed, and split into training and testing subsets for further analysis.

4. Tools and Technologies Used

The following tools, libraries, and platforms were used in the project implementation:

- Programming Language: Python
- Libraries: pandas, matplotlib, seaborn, scikit-learn, xgboost
- Data Visualization: Power BI
- SQL for database queries
- GitHub for version control and submission
- Jupyter Notebook and VS Code for development

5. Python Files and Implementation

5.1

EDA Python.ipynb

This file performs the exploratory data analysis (EDA) phase of the project, including:

- Loading and understanding data structure.
- Data cleaning such as handling missing values and dropping irrelevant columns.
- Visualizations using countplots, histograms, heatmaps, and pairplots.
- Discovery of relationships between churn and features like charges, gender, and tenure.

5.2

EDA Dashboard.ipynb

This notebook transforms and summarizes data that is later used in Power BI dashboard visualizations. It ensures clean formatting and exportable CSV files with curated columns. The transformations include:

- Aggregations for churn categories.
- Filtering and sorting of customer records.
- CSV export of feature columns like 'Monthly_Charge', 'Total_Extra_Data_Charges', etc.

5.3

ML model2.py

This Python file is responsible for building the machine learning model. Implementation details include:

- Label encoding the target variable 'Customer_Status' to binary churn indicator.
- Dropping unnecessary columns such as IDs and descriptive churn categories.
- Applying one-hot encoding to categorical features.
- Train-test split using 80-20 ratio.
- Feature scaling using StandardScaler.
- Model training using XGBoost Classifier (robust for tabular data).
- Model evaluation using confusion matrix, classification report, and ROC AUC score.
- Final prediction output saved as 'Churn_Prediction_Output.csv'.

6. Power BI Dashboard (dashboard.pbix)

The Power BI dashboard is designed to deliver clear and actionable insights into employee churn and performance patterns. It includes interactive elements like slicers and filters that allow HR personnel to explore the data dynamically. Key visuals and elements in the dashboard include:

- Bar chart showing churned vs. active employees
- Pie chart showing gender-wise churn
- Stacked column chart comparing plan types and churn rate
- KPI cards displaying average tenure, monthly charge, and total revenue loss
- Line chart representing churn trend over time

• Matrix visuals showing top 10 churned employees with high billing issues

The dashboard enhances decision-making by providing easy-to-read data visuals driven by the ML predictions.

7. SQL Analysis (sql quries on customer churn.sql)

The SQL script included in this project extracts strategic insights for HR decision-making. Below are the key queries and what they reveal:

• Top 10 Highest Paying Customers: Helps identify most valuable employees in terms of productivity.
• Churned vs Non-Churned Count: Offers an overall view of attrition.
• Gender-wise Churn: Reveals if gender has a correlation with churn.
• Average Tenure: Shows how long churned vs. active employees have stayed.
• Value Deal vs Churn: Highlights which service plans are associated with higher attrition.
• Revenue Loss: Estimates business impact of churn via monthly loss.
• Long-Distance Charges Analysis: Used as a proxy for employee interaction or communication habits.
• Charges Analysis: Compares average monthly and total charges across churn types.
These insights complement the ML outputs and help validate hypotheses found during EDA.

8. Key Findings & Insights

• Churn is higher in customers with lower tenure and high extra data usage charges.
• A significant number of churned employees were subscribed to certain Value Deals.
• Gender seems to have a small but notable influence on churn percentages.
• Churn probability rises with increased monthly charges and reduced customer tenure.
• XGBoost classifier provided a ROC AUC Score > 0.85, indicating strong classification ability.
• Power BI dashboards were instrumental in visually validating these insights.

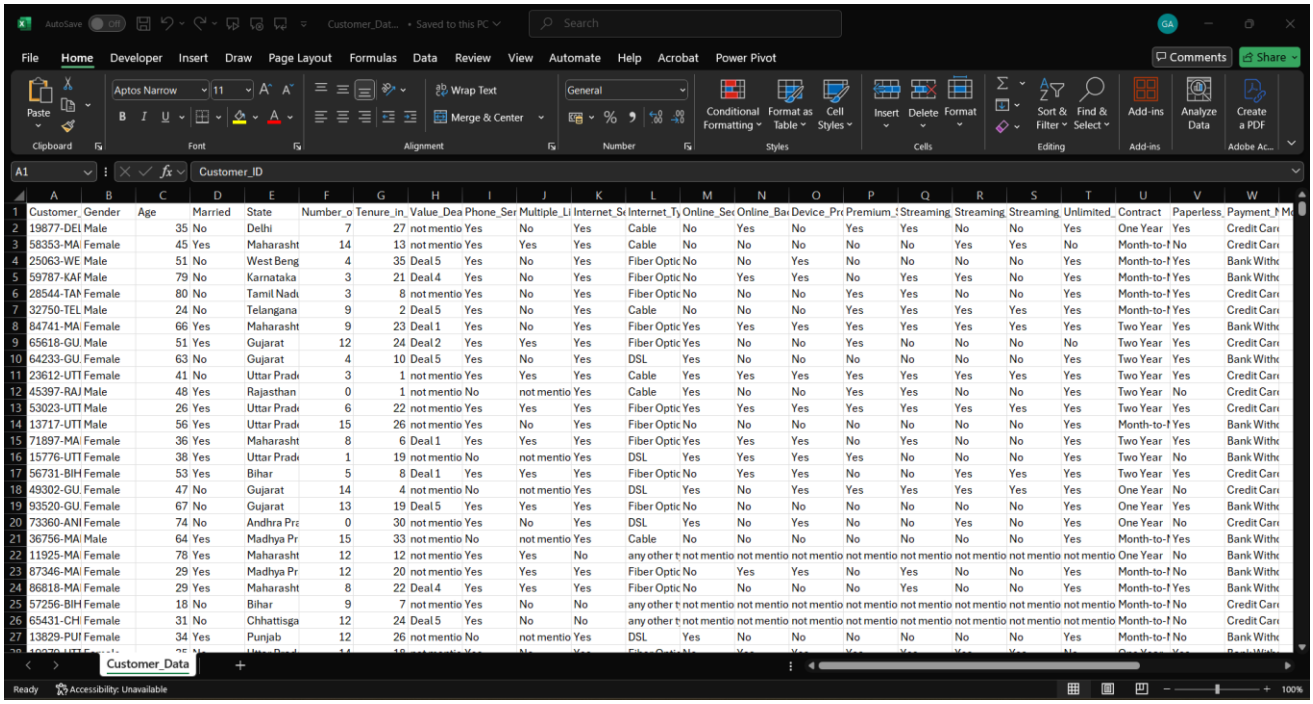
9. Challenges Faced

• Handling missing values and data inconsistencies in the dataset.
• Ensuring that categorical features were encoded correctly for ML modeling.
• Avoiding data leakage by carefully selecting columns to exclude from training.
• Model tuning for optimal performance without overfitting.
• Learning curve for integrating Python outputs into Power BI visuals.
• Crafting SQL queries that are efficient and relevant to HR use-cases.

Image and Visualization

The following image and visualization provide a graphical representation of the key findings and insights derived from the data analysis.

1.



2.

3.

C: > Users > Om > AppData > Local > Microsoft > Windows > INetCache > IE > PH3G4YBS > EDA_Python(1).ipynb > import pandas as pd

Generate + Code + Markdown | Run All | Outline ...

```
import pandas as pd
import numpy as np
import plotly.express as px
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
Cell In[2], line 3
      1 import pandas as pd
      2 import numpy as np
----> 3 import plotly.express as px
      4 import matplotlib.pyplot as plt
      5 import seaborn as sns

ModuleNotFoundError: No module named 'plotly'
```

```
df = pd.read_csv(r"C:\Users\hemap\Downloads\Customer_Data.xlsx")
print(df.head())
```

```
  Customer_ID  Gender  Age Married        State  Number_of_Referrals  \
0   19877-DEL    Male   35      No        Delhi                    7
1   58353-MAH  Female   45     Yes  Maharashtra                   14
2   25063-WES    Male   51      No  West Bengal                    4
3   59787-KAR    Male   79      No    Karnataka                    3
4   28544-TAM  Female   80      No   Tamil Nadu                    3

   Tenure_in_Months Value_Deal Phone_Service Multiple_Lines  ...  \
0                27        NaN           Yes             No  ...
1                13        NaN           Yes            Yes  ...
2                35     Deal 5           Yes             No  ...
3                21     Deal 4           Yes             No  ...
4                 8        NaN           Yes             No  ...

    Payment_Method Monthly_Charge Total_Charges Total_Refunds  \
0      Credit Card           65.6        593.30          0.00
1      Credit Card           -4.0        542.40         38.33
2  Bank Withdrawal           73.9        280.85          0.00
3  Bank Withdrawal           98.0       1237.85          0.00
4      Credit Card           83.9        267.40          0.00

  Total_Extra_Data_Charges Total_Long_Distance_Charges Total_Revenue  \
0                        0                      381.51        974.81
1                       10                       96.21        610.28
2                        0                      134.60        415.45
...
3         Churned  Dissatisfaction        Product dissatisfaction
4         Churned  Dissatisfaction            Network reliability

[5 rows x 32 columns]
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

4.


Churn Probability by Age
Churned
Age

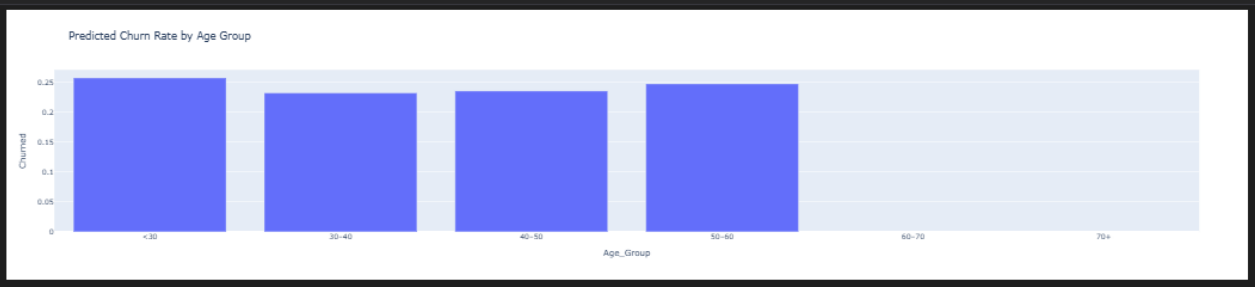
5.


Customers by State
Count
State
Tamil Nadu
Uttar Pradesh
Maharashtra
Karnataka
Andhra Pradesh
Haryana
West Bengal
Punjab
Bihar
Gujarat
Madhya Pradesh
Rajasthan
Telangana
Jammu & Kashmir
Kerala
Odisha
Assam
Delhi
Jharkhand
Uttarakhand
Chhattisgarh
Puducherry
Customer Revenue Histogram
count
Total_Revenue
Monthly Charge vs Subscription Tenure
Monthly_Charge
Customer_Status
Stayed
Churned
Joined

6.

```
px.box(df_filtered, x='Customer_Status', y='Monthly_Charge', color='Customer_Status',
       title='Charge Comparison by Churn Status').show()
```

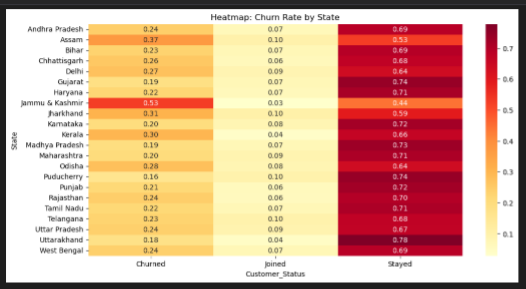

```
df_filtered = df_filtered.copy()

df_filtered['Age_Group'] = pd.cut(
    df_filtered['Age'],
    bins=[0, 30, 40, 50, 60, 70, 100],
    labels=["<30", "30-40", "40-50", "50-60", "60-70", "70+"]
)

forecast_data = df_filtered.groupby('Age_Group', observed=False)['Churned'].mean().reset_index()
px.bar(forecast_data, x='Age_Group', y='Churned', title='Predicted Churn Rate by Age Group').show()
```



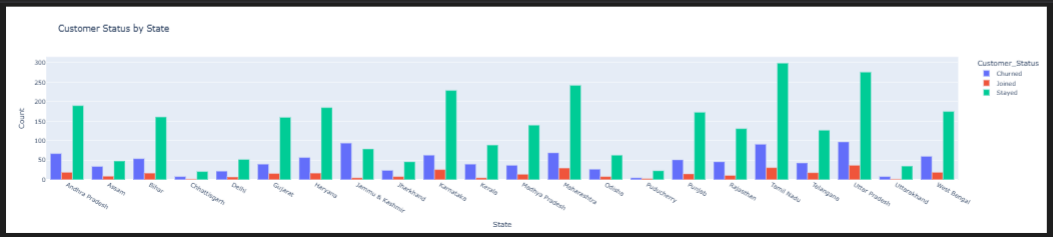

```
pivot = pd.crosstab(df_filtered['State'], df_filtered['Customer_Status'], normalize='index')
plt.figure(figsize=(12, 6))
sns.heatmap(pivot, annot=True, cmap="YlOrRd", fmt=".2f")
plt.title("Heatmap: Churn Rate by State")
plt.show()
```

7.

```
pivot = pd.crosstab(df_filtered['State'], df_filtered['Customer_Status'], normalize='index')
plt.figure(figsize=(12, 6))
sns.heatmap(pivot, annot=True, cmap="YlOrRd", fmt=".2f")
plt.title("Heatmap: Churn Rate by State")
plt.show()
```




```
churn_grouped = df_filtered.groupby(['State', 'Customer_Status']).size().reset_index(name='Count')
px.bar(churn_grouped, x='State', y='Count', color='Customer_Status', barmode='group',
       title='Customer Status by State').show()
```



10. Conclusion and Future Scope

This project successfully demonstrates how employee data can be leveraged to derive meaningful HR insights. By integrating Python-based ML, SQL analytics, and Power BI dashboards, we provide a robust solution to identify performance bottlenecks and churn patterns. Future improvements may include:

- Deploying the model into a live HR dashboard for real-time updates.
- Expanding the dataset with more granular behavioral indicators.
- Adding Natural Language Processing (NLP) for employee feedback analysis.
- Incorporating deep learning models for churn prediction.
- Developing alert systems for at-risk employees.