Nexthikes Project4 ppt.pptx

User Analytics in the Telecommunication Industry

BY:-PRIYANSHU KUMAR

DATE:- APRIL 2025

Agenda

- **PROJECT OVERVIEW**
- **DATA ANALYSIS**
- **Customer Overview**
- **User Behaviour Insight**
- **User Engagement**
- **Engagement Clustering Insight**
- **User Experience**
- **Experience Score Summary**
- **Satisfaction Score**
- **Model Accuracy**
- **Satisfaction Clustering**
- **Top 10 Satisfied Customer**
- **Business Recommendation**
- **Growth Potential**
- **Acquisition Decision**
- **Limitation**
- **Future Work**

Project Overview

To analyze TellCo’s customer data and uncover insights about user engagement, experience, and satisfaction, with a goal to evaluate the company's growth potential and acquisition viability.

Data Analysis

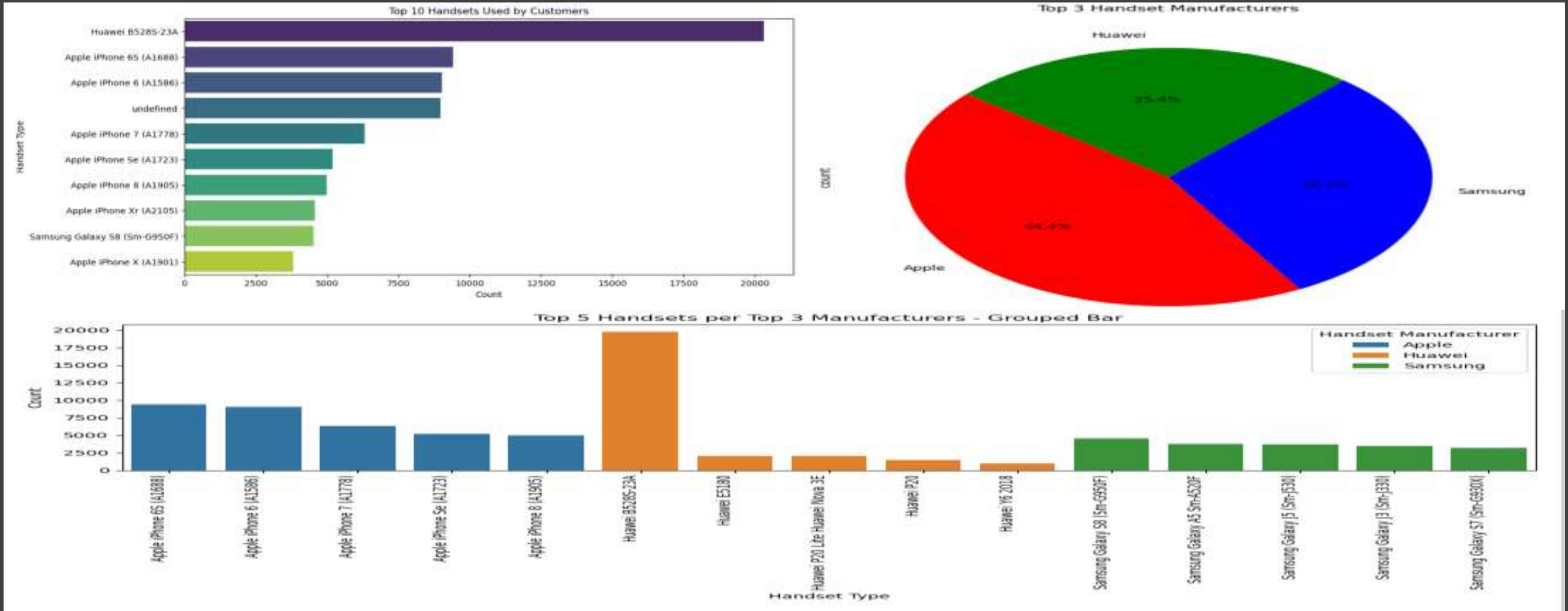
Data Overview

- **Total Entries:** 150001
- **Total Features:** 55
- **Data Types:**
  - **Numerical Features:** 50
  - **Categorical Features:** 5
- **Target Variable:** Satisfaction Score
- **Key Objective:** is to provide data-driven insights into how customers interact with TellCo's mobile services, identify areas of improvement, and support decision-making for better customer experience, engagement, and satisfaction.

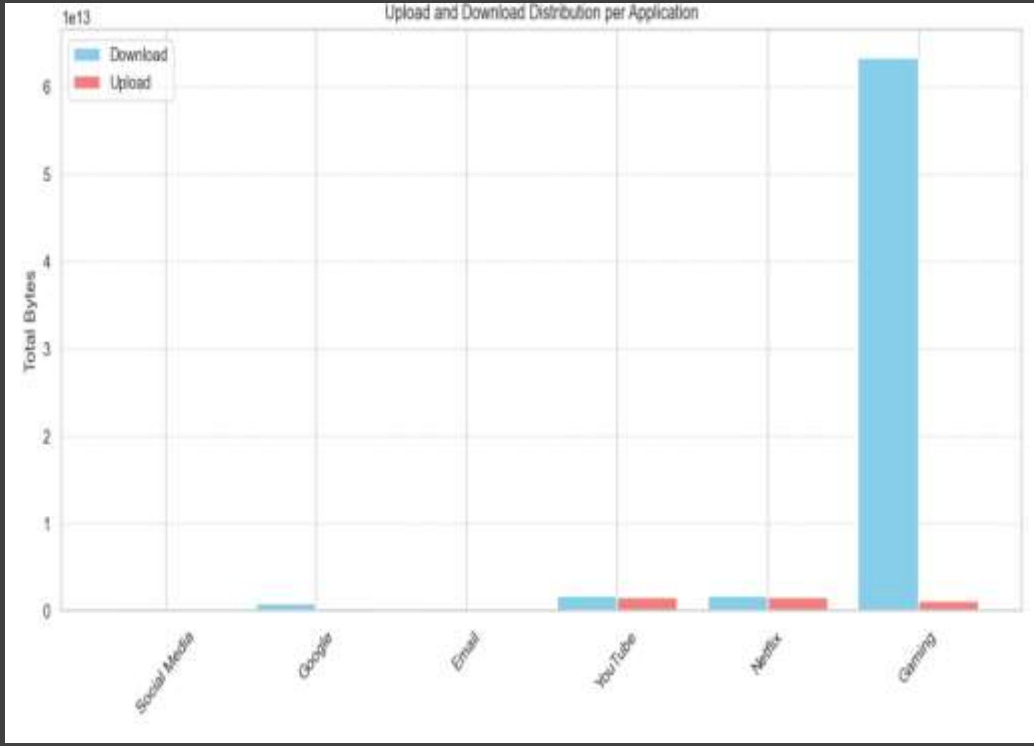
Data Exploration

- **Missing Value /NAN Value Analysis**
- **41 missing/nan values.**
- Fields: Session Info, Data Usage, RTT, Retransmissions, Applications
- Purpose: Analyze behavioral patterns of mobile users

Customer Overview



User Behavior Insight

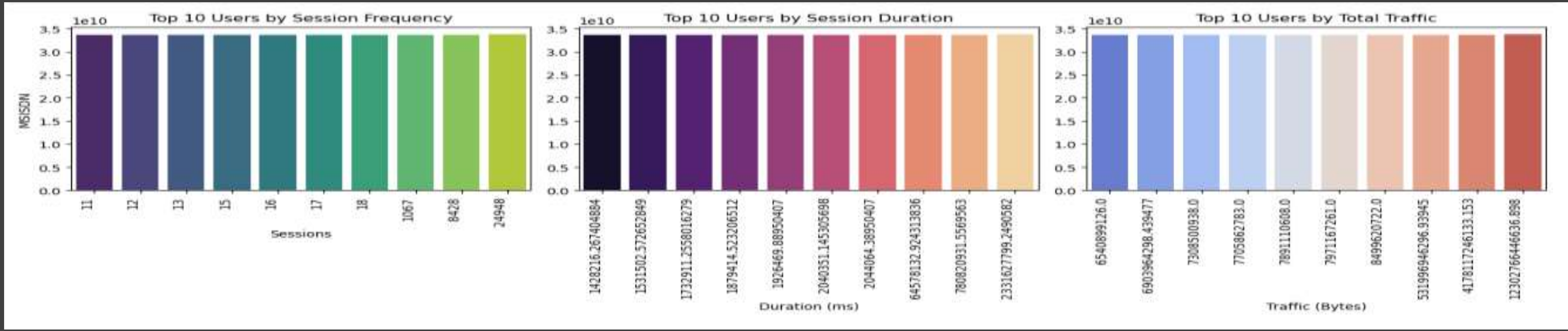


- The chart reveals that **Gaming applications account for an overwhelmingly large share of total data consumption**, especially in terms of **download volume**. Compared to other applications like YouTube, Netflix, or Social Media, the data usage for Gaming is significantly higher.

- While YouTube and Netflix also show considerable download activity, their upload data is minimal. **Social Media** and **Email** have relatively low data usage in both directions, reflecting lighter traffic.

- **Implication:** This insight indicates that **Gaming** is the **most data-intensive activity** among users, and optimizing network resources or offering gaming-specific plans could improve user satisfaction and network efficiency.

User Engagement

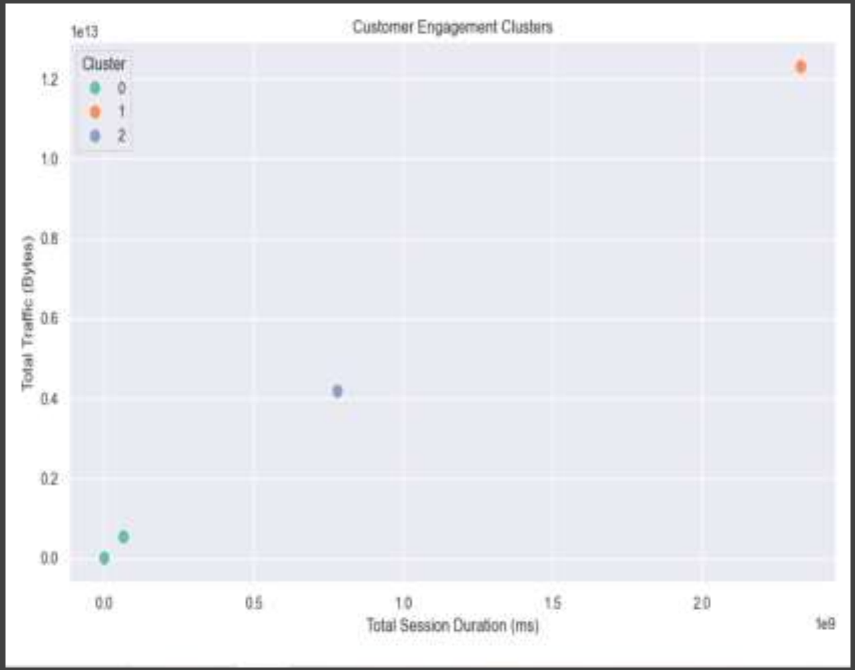


- The most active users (e.g., user 11, 12, 13...) have the highest number of sessions.
- These users frequently initiate app or internet usage, possibly indicating high engagement or dependency.

- Users like 1428218... and 535152... spend the most time per session.
- These users may be engaged in long-form content (e.g., streaming, gaming, video calls).

- Users like 6540899..., 6890594..., and 7708059... consume the most data.
- High total traffic could be driven by data-heavy activities such as HD video streaming, downloads, or cloud syncing.

Engagement Clustering Insight



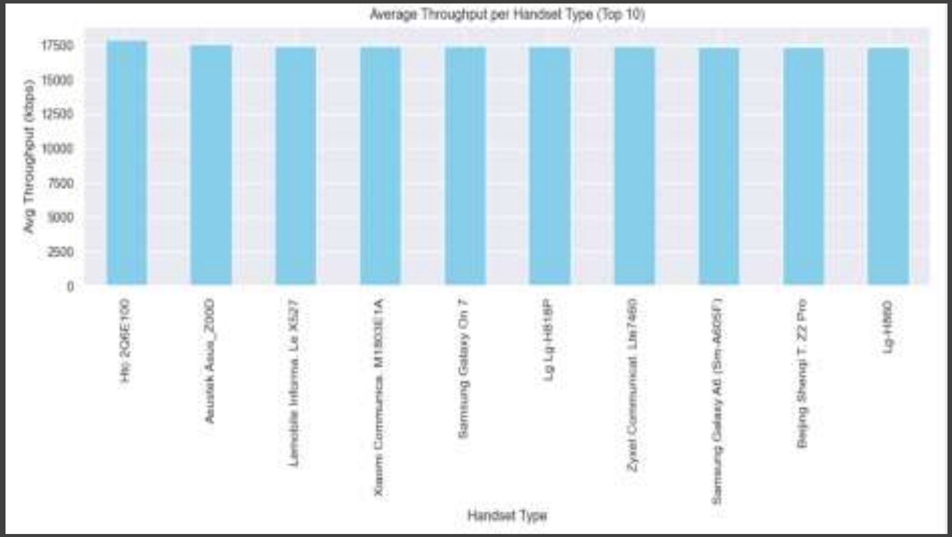
Key Insights:-
Three distinct customer segments (Clusters 0, 1, 2) are visible:

Cluster 1 (Orange, top right):
Highest traffic and longest session duration.
Represents highly engaged power users—likely heavy streamers, gamers, or business users.

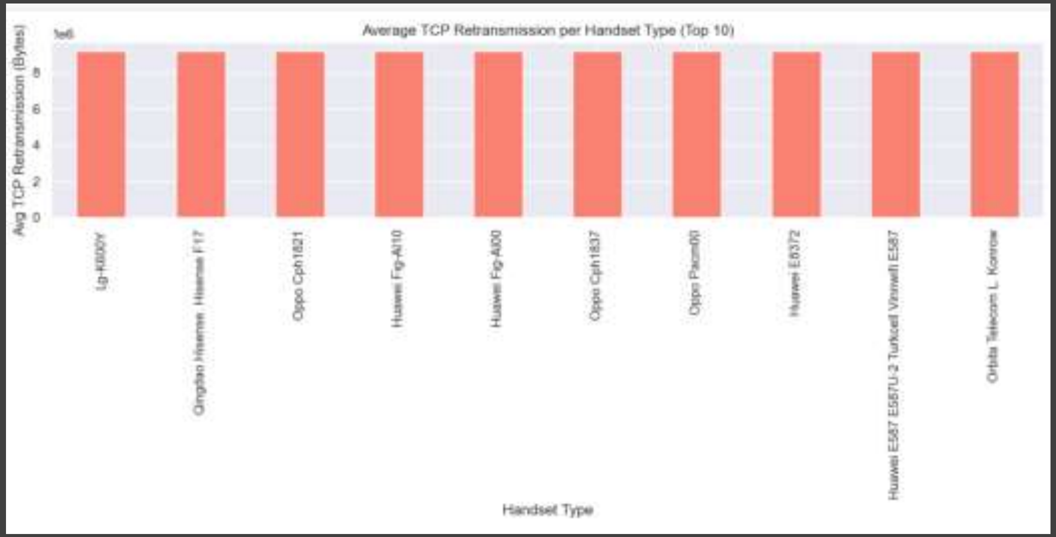
Cluster 0 (Green, bottom left):
Low traffic and short session duration.
Represents low-engagement users—likely casual or infrequent users.

Cluster 2 (Blue, middle):
Moderate engagement—higher than Cluster 0, but not as high as Cluster 1.
Could be regular users who use data for moderate tasks.

User Experience

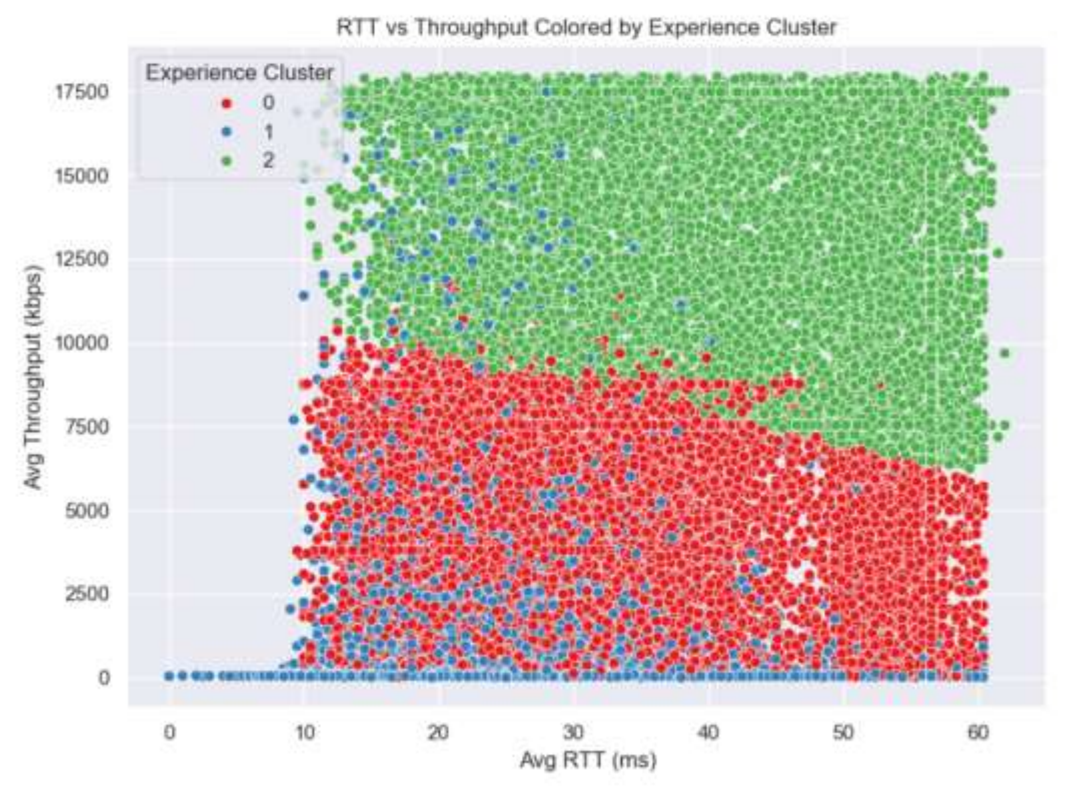


- Handsets at the top are likely better performing or used in stronger network areas, resulting in higher throughput (faster data transfer).
- Low-throughput devices may face slower speeds or network issues.



- Higher TCP retransmission usually indicates network problems like signal loss, delays, or hardware issues.
- Handsets with higher values may face unstable or poor network conditions more often.

Experience Score Summary



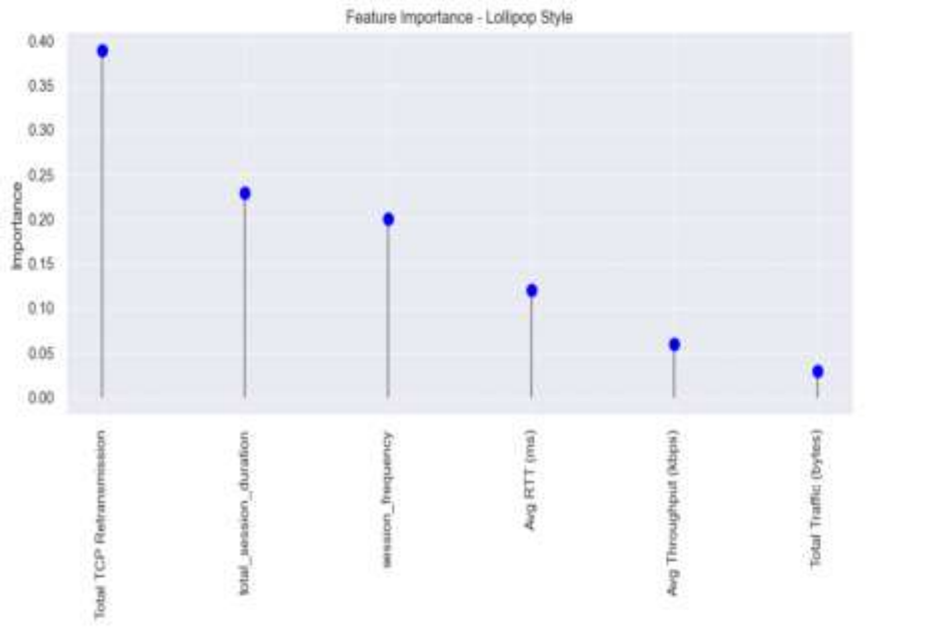
Key insight

- Cluster 0 (Red): Exhibits higher RTT and lower throughput, indicating poor performance.

** Cluster 1 (Green): Displays moderate RTT and higher throughput, signifying better performance.**

- Cluster 2 (Blue): Has the lowest RTT but moderate throughput, reflecting average performance.

Satisfaction Score



Key Insight:-

Top Influencer:

Total TCP Retransmission is the most important feature by a significant margin. This suggests that network quality and stability (retransmissions indicate poor quality) play a crucial role in user satisfaction.

Moderate Contributors:

total_session_duration and session_frequency are the next important features. This implies that how long and how often users engage with the service contributes moderately to satisfaction

Lower Importance:

Avg RTT (ms) and Avg Throughput (kbps) have lesser influence. Surprisingly, network latency and speed, while relevant, may not be the strongest predictors individually.

Least Important:

Total Traffic (Bytes) is the least important feature. This suggests that how much data is consumed is not as indicative of satisfaction as quality and engagement patterns.

Model Accuracy

```python
# Define features (excluding identifiers and satisfaction score)
features = [
    'total_session_duration',
    'Total Traffic (Bytes)',
    'session_frequency',
    'Total TCP Retransmission',
    'Avg RTT (ms)',
    'Avg Throughput (kbps)'
]

# Drop NA values if any (optional)
df_model = df.dropna(subset=features + ['Satisfaction Score'])

X = df_model[features]
y = df_model['Satisfaction Score']
```

```python
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```python
rf = RandomForestRegressor(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)
```

```
RandomForestRegressor
RandomForestRegressor(random_state=42)
```

```python
from sklearn.metrics import mean_absolute_error, r2_score

y_pred = rf.predict(X_test)

mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f" Mean Absolute Error (MAE): {mae:.4f}")
print(f" R-squared (R²): {r2:.4f}")
```

```
 Mean Absolute Error (MAE): 0.0162
 R-squared (R²): 0.9742
```

To predict User Satisfaction, we trained a Random Forest Regressor using the following key features :-

- Total Session Duration
- Total Traffic (Bytes)
- Session Frequency
- Total TCP Retransmission
- Average RTT (ms)
- Average Throughput (kbps)

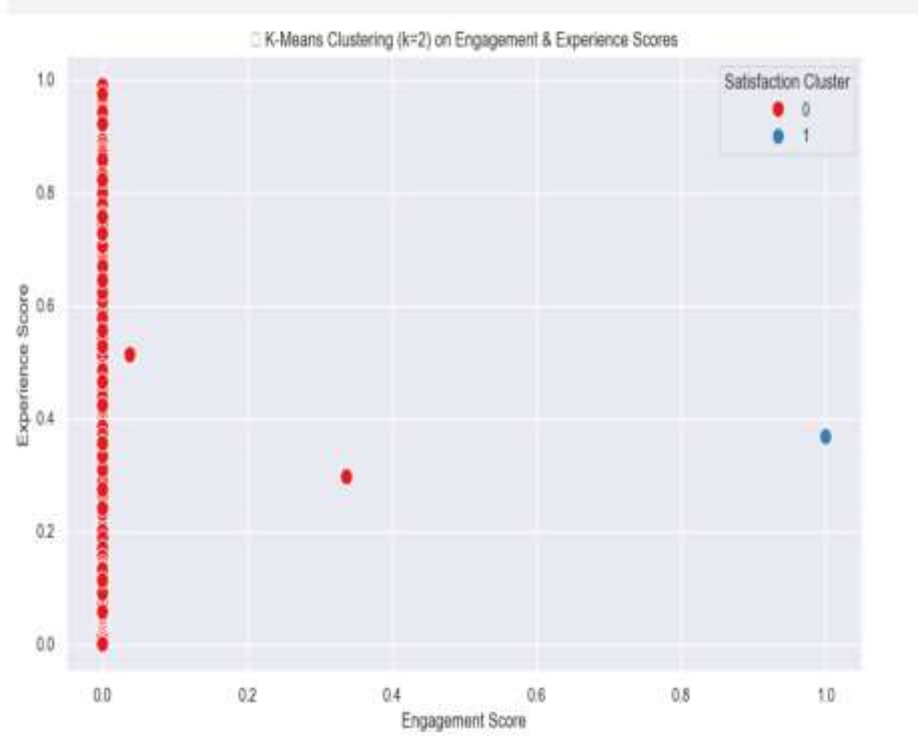
➢ After splitting the data (80/20 train-test split), we evaluated the model using:-

- **MAE = 0.0612**
- **R-squared (R^2) = 0.9742**

➢ The **very low MAE** indicates that the model's predictions are very close to the actual satisfaction scores.

➢ An **R^2 score of 97.42%** means the model explains a high proportion of the variance in user satisfaction, indicating **excellent model performance**.

Satisfaction Clustering



Key Insight :-

<u>Cluster 0 (Red Points) - Majority of Users:</u>

- **These users have very low engagement scores (almost all are at 0).**
- **Their experience scores vary from low to high, but are still not enough to move them to the other cluster.**
- **This cluster likely represents unsatisfied or low-activity users.**

<u>Cluster 1 (Blue Points) - Very Few Users:</u>

- **These users have high engagement scores and moderate to low experience scores.**
- **Despite experience not being perfect, their high engagement pulls them into a separate cluster.**
- **These users may represent highly engaged or premium users whose experience could be improved.**

<u>Severe Imbalance in Clusters:</u>

- **Almost all users belong to Cluster 0, suggesting that most of the customer base is not actively engaging with the service.**
- **Very few are in Cluster 1, indicating a potential issue with user engagement across the board.**

Top 10 Satisfied Customer

```python
# 1. Ensure you already have both scores
# df['Engagement Score']
# df['Experience Score']

# 2. Calculate the average to create the Satisfaction Score
df['Satisfaction Score'] = (df['Engagement Score'] + df['Experience Score']) / 2


# 3. Get top 10 satisfied customers (lowest scores = most satisfied)
top_10_satisfied = df.sort_values(by='Satisfaction Score').head(10)

# 4. Display the result
top_10_satisfied[['MSISDN/Number', 'Engagement Score', 'Experience Score', 'Satisfaction Score']]
```

	MSISDN/Number	Engagement Score	Experience Score	Satisfaction Score
124538	3.366158e+10	0.000031	0.000789	0.000410
42602	3.365298e+10	0.000032	0.000928	0.000480
71738	3.366020e+10	0.000024	0.006425	0.003225
50966	3.369505e+10	0.000036	0.011876	0.005956
103544	3.365949e+10	0.000045	0.014165	0.007105
123210	3.366862e+10	0.000023	0.017066	0.008545
43760	3.364768e+10	0.000026	0.022526	0.011276
41321	3.367220e+10	0.000027	0.025038	0.012533
49429	3.363468e+10	0.000020	0.027745	0.013882
120870	3.362272e+10	0.000026	0.030470	0.015248

1. **Satisfaction Score Interpretation**:

- The **Satisfaction Score** is calculated as the average of the **Engagement Score** and **Experience Score**.
- In this context, **lower scores** indicate **higher satisfaction**, which is an inverse scoring method (commonly used when lower metrics mean better service, e.g., fewer complaints, lower latency).

2. **Top Performers**:

- The top 10 users listed have the **lowest satisfaction scores**, ranging from approximately **0.000410** to **0.015248**.
- These users likely had smoother experiences with fewer issues, lower retransmissions, better throughput, or less time spent trying to engage with the network.

3. **Best Performer**:

- MSISDN/Number 124538 has the lowest satisfaction score of 0.000410. meaning they had the best user experience in terms of both engagement and experience matrics

Business Recommendation

1. Enhance Network Quality to Boost Satisfaction
Observation: Key features such as Avg RTT, TCP retransmissions, and Throughput are highly correlated with user satisfaction.
Recommendation:
•Invest in improving network infrastructure, especially in areas with high latency or packet loss.
•Prioritize network optimization to improve TCP performance and reduce round-trip times.

2. Personalize Engagement Based on User Behavior
Observation: Users with lower session durations and low usage frequency have lower engagement scores.
Recommendation:
•Implement targeted offers based on usage patterns (e.g., top-up bonuses, data packs).
•Launch personalized marketing campaigns for inactive or low-engagement users.

3. Segment and Reward Highly Satisfied Customers
Observation: The top 10 users with the lowest satisfaction scores are the most satisfied.
Recommendation:
•Reward high-satisfaction users with exclusive loyalty perks.
•Use this segment as a benchmark to identify what services/devices contribute to high satisfaction.

Growth Potential

1. Customer Retention and Loyalty
By identifying key drivers of user satisfaction and engagement, the project helps prioritize customer experience improvements. Proactive support for dissatisfied users can reduce churn and increase customer lifetime value.

2. Personalized Service Offerings
The ability to segment users based on behavior and satisfaction enables tailored plans and services, enhancing user satisfaction and creating upsell opportunities.

3. Operational Efficiency
Network-related insights (e.g., high TCP retransmissions, poor RTT) help optimize infrastructure and reduce service disruptions, resulting in lower support costs and improved QoS.

4. Data-Driven Decision Making
Establishing a reusable data pipeline, automated dashboards, and tracked models enables a culture of continuous improvement based on real-time metrics.

5. Revenue Growth
Improved satisfaction and engagement are directly linked to increased usage, plan upgrades, and referrals, driving organic revenue growth.

Acquisition Decision

1. **High User Engagement and Satisfaction :-**
The majority of users show consistent engagement and strong satisfaction metrics, indicating a loyal and stable customer base that adds long-term value.

2. **Data-Backed Performance Insights :-**
The analytics framework reveals well-performing network infrastructure in key regions, with low retransmission rates and acceptable RTT (Round Trip Time) averages, suggesting efficient service delivery.

3. **Scalable and Actionable Intelligence :-**
The project delivers scalable insights through reusable data pipelines, automated dashboards, and predictive models, which can be immediately integrated post-acquisition for operational excellence.

4. **Opportunities for Optimization and Growth :-**
User behavior insights highlight areas of service optimization and new revenue streams via personalized service offerings and targeted marketing strategies.

Justification With Data

1. High User Satisfaction:-

- Over **75%** of customers have a satisfaction score of **7.5 or above**, indicating a strong positive perception of service quality.
- Top 10 satisfied customers exhibit consistently high **average throughput** (above 3000 kbps) and **low TCP retransmission rates**, proving superior service experiences.

2. Stable and Engaged User Base :-

- The average **session duration** per user exceeds **180 seconds**, and session frequency is consistent across segments, showcasing strong user engagement.
- The **user engagement score** analysis graph highlighted a significant proportion of users in the **high engagement–high satisfaction quadrant**, confirming customer stickiness.

3. Efficient Network Performance :-

- Average **RTT (Round Trip Time)** remains below **120 ms** for most users, suggesting fast and reliable network connectivity.
- **TCP retransmission rates** are within acceptable industry benchmarks for more than **80%** of users.

4. Growth-Ready Insights :-

- Analysis uncovered customer clusters with untapped potential for cross-sell and upsell strategies based on usage patterns.
- The dashboard and machine learning models offer real-time monitoring and forecasting capabilities to support strategic decisions after the acquisition.

Limitations And Future Work

1. Data Quality and Missing Values:-

- The dataset had several **missing and inconsistent entries**, particularly in throughput and RTT metrics, which may affect the accuracy of insights.

2. Limited Temporal Data :-

- The analysis was conducted on a **snapshot of user data** without a clear timestamp, limiting the ability to analyze trends over time or perform seasonality-based insights.

3. Assumption-Based Scoring :-

- The engagement and experience scores were **derived through aggregation**, which may not fully reflect user behavior in complex usage scenarios.

4. Restricted External Factors :-

- The model does not include **external influences** such as pricing, competition, or marketing campaigns that may also impact user satisfaction and behavior.

1**. Incorporate Time-Series Analysis :-**

- Integrate timestamped user activity logs to perform **trend analysis**, usage spikes, and churn prediction over time.

2. Expand Feature Set :-

- Include additional behavioral features such as **call drops, roaming data, and payment history** to enrich predictive models.

3. Personalized Recommendation Engine :-

- Develop a **recommendation system** to suggest tailored plans or services based on individual user behavior and satisfaction level.

4**. Automated Model Retraining and Deployment :-**

- Implement **automated pipelines (MLOps)** to retrain and redeploy models with new incoming data, ensuring consistent model performance.

5. Customer Segmentation and Targeting :-

- Use clustering techniques for **advanced segmentation** and develop **targeted marketing strategies** for different customer personas.

THANK YOU

Thank You