Telecom Market Share Churn Prediction

The project aims to develop a model that predicts which operator-circle combinations are at risk of subscriber decline (market churn), enabling proactive retention and competitive strategies in the market.

Presented by

Team D

Date

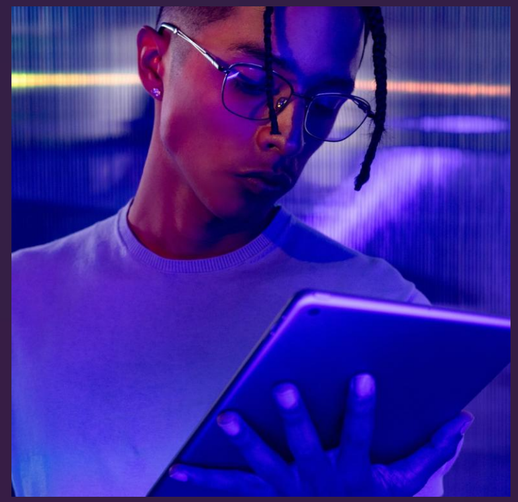
Oct 12th, 2025

Evoastra

Our Team

Aditya Sakinal

Team Lead



Naga Sri Durga Mallesh Tanna

Co Team Lead 1

Sangeeta Prajapati

Co Team Lead 2

Team members

M.K Aysha Meharin
Monu Pal
Geethanjali Banda
Rikshith Bommena
Ankit Thummar
Dayakar Uppuluri
Venkata Karthik Dhulipudi
Muheeb Mohammad
Santhoshini Gundreddy
Kranthi Kumar Gowni
Thirumala Devi Kommalapati

Introduction

The Proposed System successfully developed a Time-Series Telecom Churn Prediction System to proactively identify operator-month segments at risk of subscriber decline (market churn), enabling targeted, high-ROI retention efforts. Leveraging granular multi-year data (2009–2025(till April)), We engineered sophisticated velocity-of-change trend features and competitive metrics (HHI). After rigorous model comparison across ML and DL architectures, Random Forest model was selected for its superior performance, achieving a 0.9979 F1-Score. The system is deployed via a Streamlit Dashboard, delivering a projected 1,256.15% ROI from targeted intervention, along with SHAP-based feature importance and critical 2025–2026 revenue loss forecasts for actionable competitive intelligence.

About the Project

The Proposed System focuses on building a **machine learning–based market churn prediction system** using **TRAI telecom data**. The model analyzes multi-year subscriber records across different operators and geographic circles to identify **where and when subscriber losses are most likely to occur**.

In the competitive telecom landscape, even small declines in key circles can trigger significant revenue loss and reduced market presence. This project transforms churn prediction from a reactive process to a proactive strategy, giving telecom providers the power to anticipate subscriber loss before it happens.

Unlike conventional customer-level churn models, this approach analyzes market-wide dynamics—capturing regional, temporal, and operator-specific trends. The result is a robust analytical framework that guides marketing investment, network expansion, and strategic planning with data-driven foresight.

By blending machine learning intelligence with business strategy, the solution helps the client strengthen market positioning, retain customers, and stay ahead in a rapidly shifting telecom environment.

Project Objective

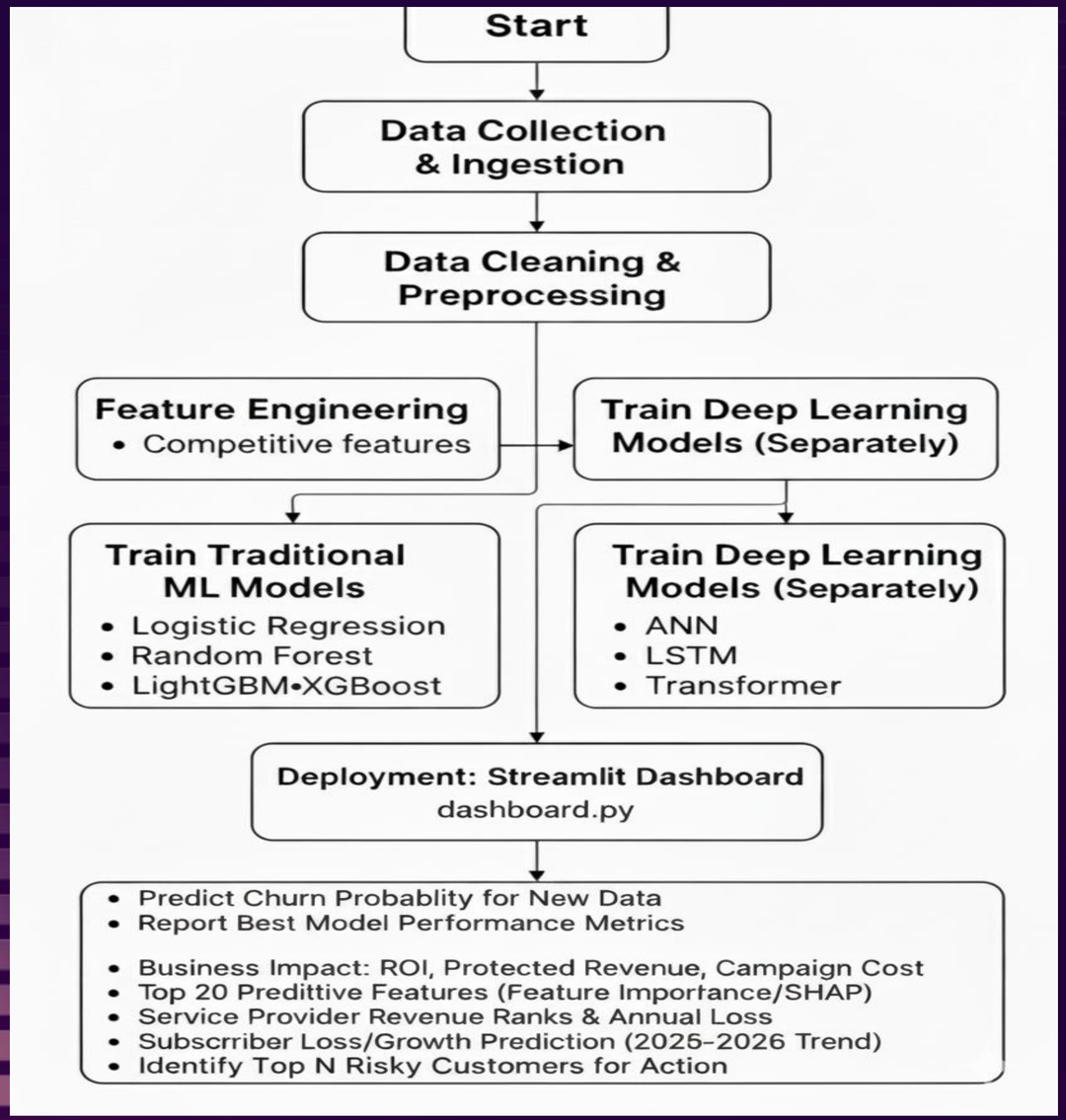
To design and implement a predictive analytics system that:

- Forecasts market-level subscriber churn across different operators, circles, and time periods using historical TRAI telecom data.
- Identifies high-risk regions and operator segments most likely to experience subscriber decline, enabling early action.
- Facilitates data-driven decisions that enhance competitive positioning, profitability, and market sustainability.
- Integrates machine learning and deep learning models (XGBoost, LightGBM, Neural Networks) for precise, explainable, and scalable predictions.
- Transforms static data into dynamic intelligence, providing telecom leaders with real-time visibility into market health.
- Identifies high-risk circles and potential loss zones and Enables data-driven decisions that strengthen competitive advantage.

Tools & Technologies Used

- **Programming & IDE:** Python, VS Code
- **Data Handling & Storage:** Pandas, NumPy, CSV / Excel
- **Machine Learning:** Scikit-learn, XGBoost, LightGBM
- **Deep Learning:**, TensorFlow, Transformers, LSTM
- **Model Optimization & Deployment:** Joblib, Optuna, Streamlit
- **Visualization & Insights:** Matplotlib, Seaborn, SHAP, Power B
- **Configuration & Utilities:** JSON, OS
- **Presentation:** Microsoft PowerPoint

Project Workflow



Dataset Cleaning

To ensure the dataset was accurate, consistent, and ready for analysis, several cleaning steps were performed:

- **Data Quality/Filtering:**
  Action: Removed **85 duplicate rows** and filtered out 22,654 rows where value=0.
  Source & Outcome: Retained ≈48,000 active subscriber records, eliminating unusable noise.

- **Data Type Standardization:**
  Action: Corrected the critical value column from Object (string) type by imputing missing/non-numeric values with 0 and casting to int.
  Source & Outcome: Established a **reliable numeric metric** for subscriber count.

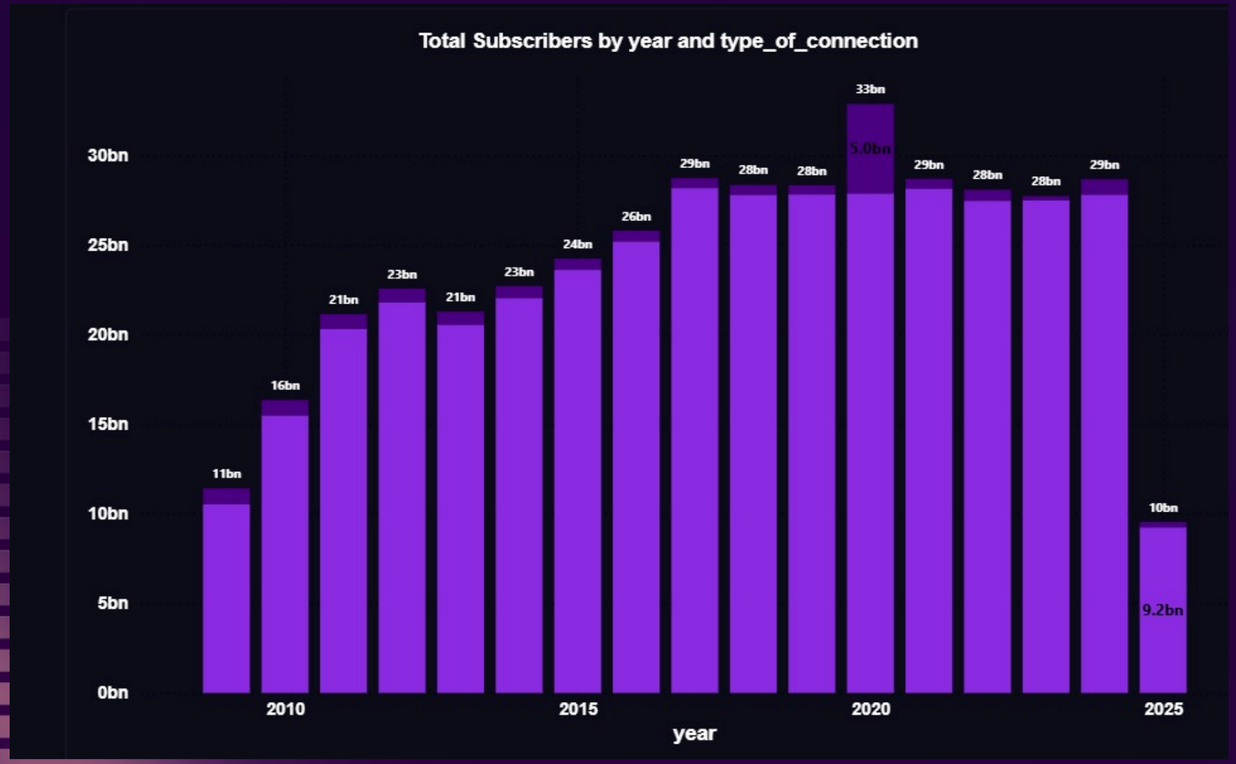
- **Categorical Consistency:**
  Action: Standardized 42 service_providers and 50 circle names by trimming whitespace and converted them to the **category** data type.
  Source & Outcome: Ensured cleaner inputs for encoding and optimized memory usage.

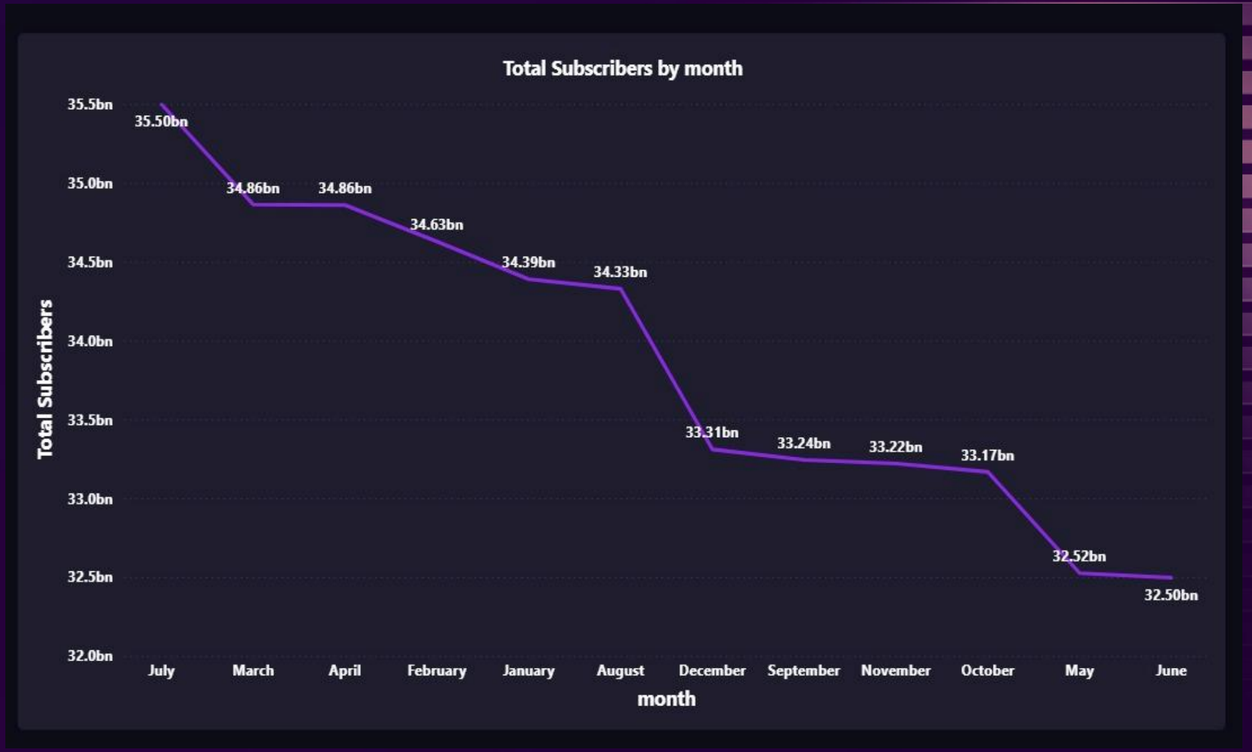
- **Final Data Partition:**
  Action: The cleaned dataset (70,728 rows total) was partitioned for the ML workflow.
  Source & Outcome: Created specific 10% (df1), 10% (df2), and 80% (df3) subsets for training and validation.

Dataset Overview

	year	month	circle	type_of_connection	service_provider	value	unit	notes
0	2018	March	Assam	wireless	BSNL	1964401	value in absolute number	NaN
1	2018	March	Bihar	wireless	BSNL	4430024	value in absolute number	NaN
2	2018	March	Delhi	wireless	BSNL	0	value in absolute number	NaN
3	2018	March	Gujarat	wireless	BSNL	5810246	value in absolute number	NaN
4	2018	March	Haryana	wireless	BSNL	4476401	value in absolute number	NaN

Data Visualizations

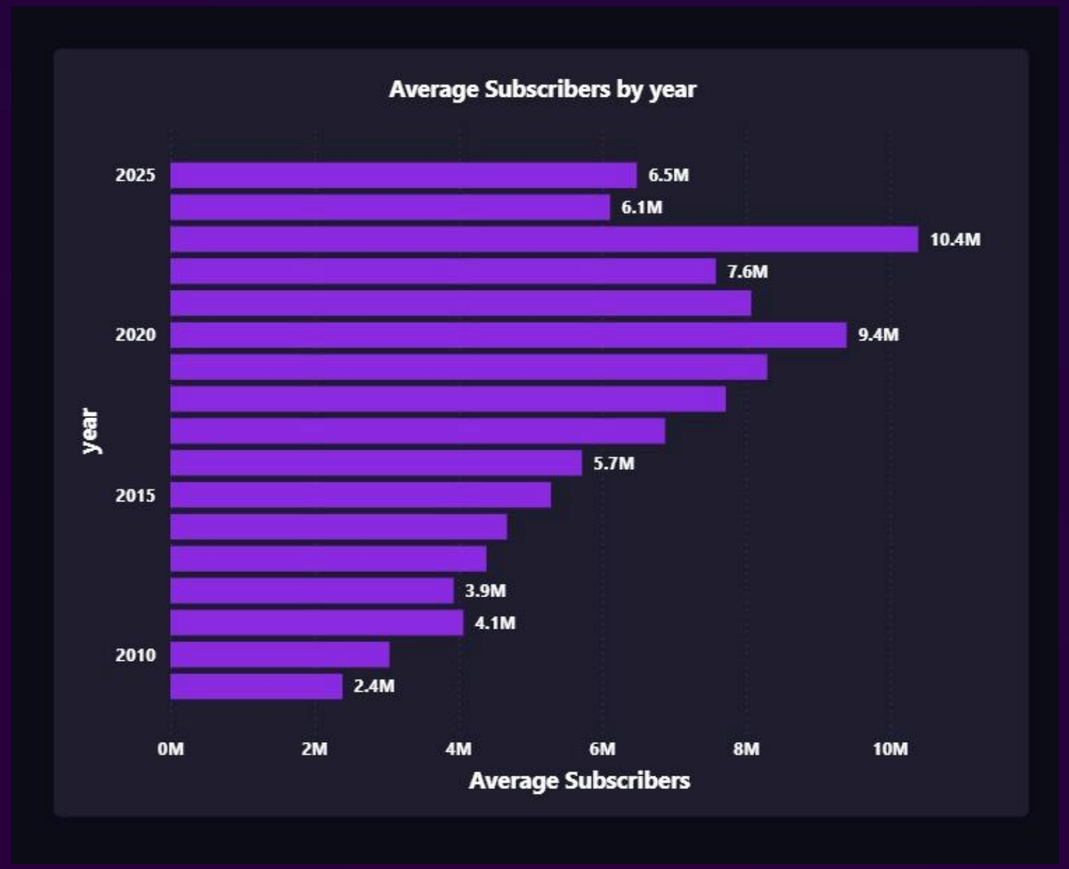




Best Provider

Bharti Airtel

First service_provider



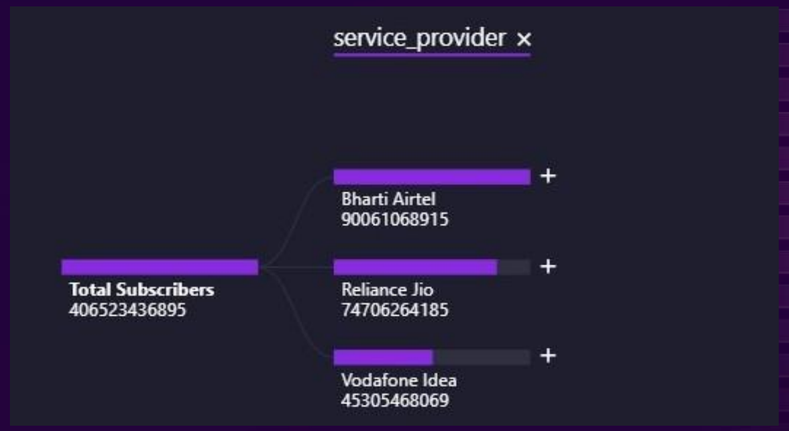
circle	service_provider	Total Subscribers
All India	Vodafone Idea	22652734039
Andhra Pradesh	Vodafone Idea	1209548037
Assam	Vodafone Idea	248461974
Bihar	Vodafone Idea	968553190
Delhi	Vodafone Idea	1375889407
Gujarat	Vodafone Idea	1997142147
Haryana	Vodafone Idea	656110615
Himachal Pradesh	Vodafone Idea	55619860
Jammu And Kashmir	Vodafone Idea	44205020
Karnataka	Vodafone Idea	756402254
Kerala	Vodafone Idea	1316940240
Kolkata	Vodafone Idea	538082498
Madhya Pradesh	Vodafone Idea	1680754147
Maharashtra	Vodafone Idea	2504520509
Mumbai	Vodafone Idea	977568051
North East	Vodafone Idea	106851364
Odisha	Vodafone Idea	201837069
Punjab	Vodafone Idea	676203554
Rajasthan	Vodafone Idea	989920635
Tamil Nadu	Vodafone Idea	1508815025
Uttar Pradesh (East)	Vodafone Idea	1848333349
Total		406523436895

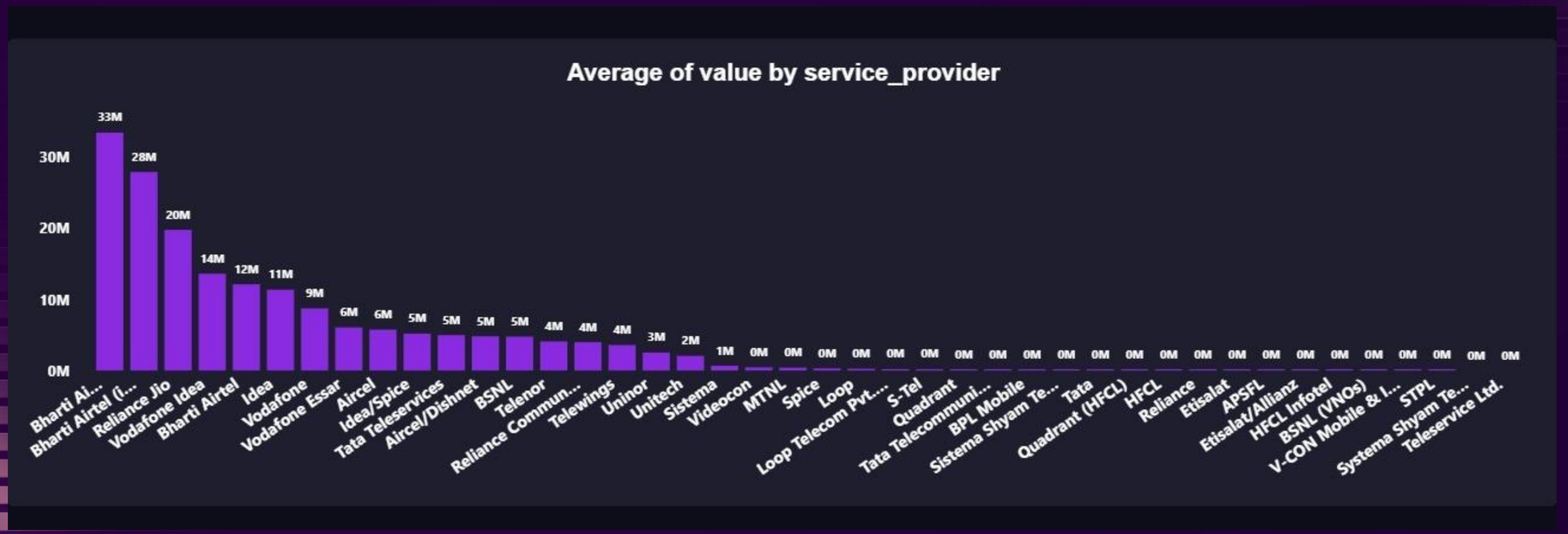
Sum of value by service_provider

45305468069

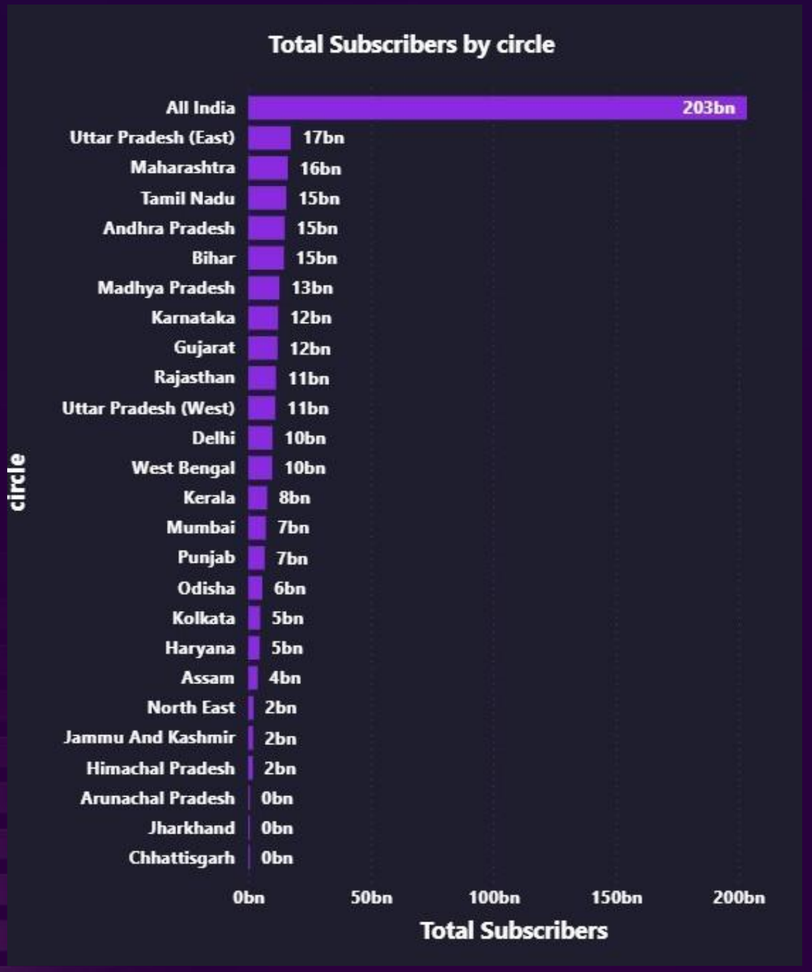
Table title

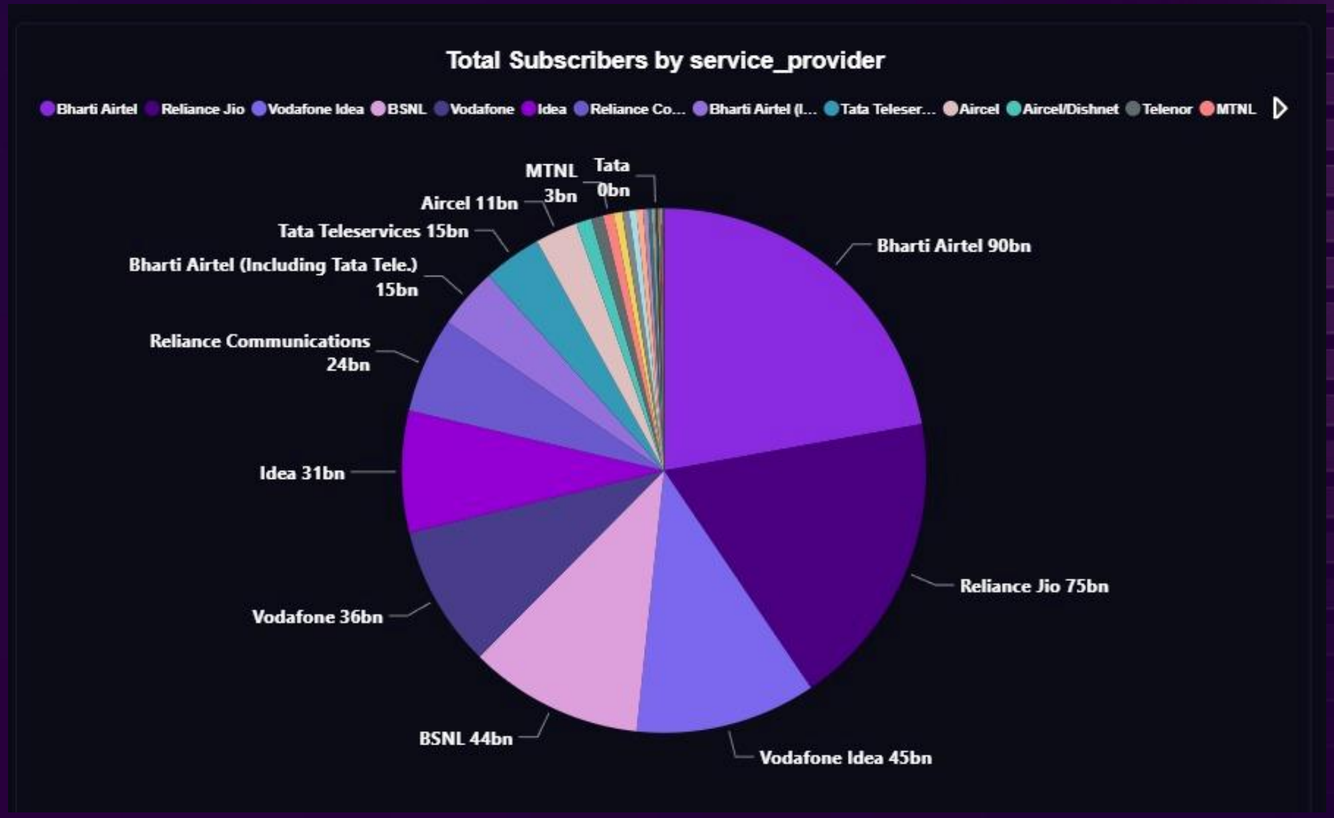
circle	service_provider	Sum of value
All India	Vodafone Idea	22652734039
All India	Vodafone	17767627960
All India	Tata Teleservices	7310866592
All India	Reliance Jio	37234173994
All India	Reliance Communications	12185857613
All India	Idea	15298427740
All India	BSNL	20679066592
All India	Bharti Airtel (Including Tata Tele.)	7699102970
All India	Bharti Airtel	44791994661
All India	Aircel	5307679253
Total		**190927531414**

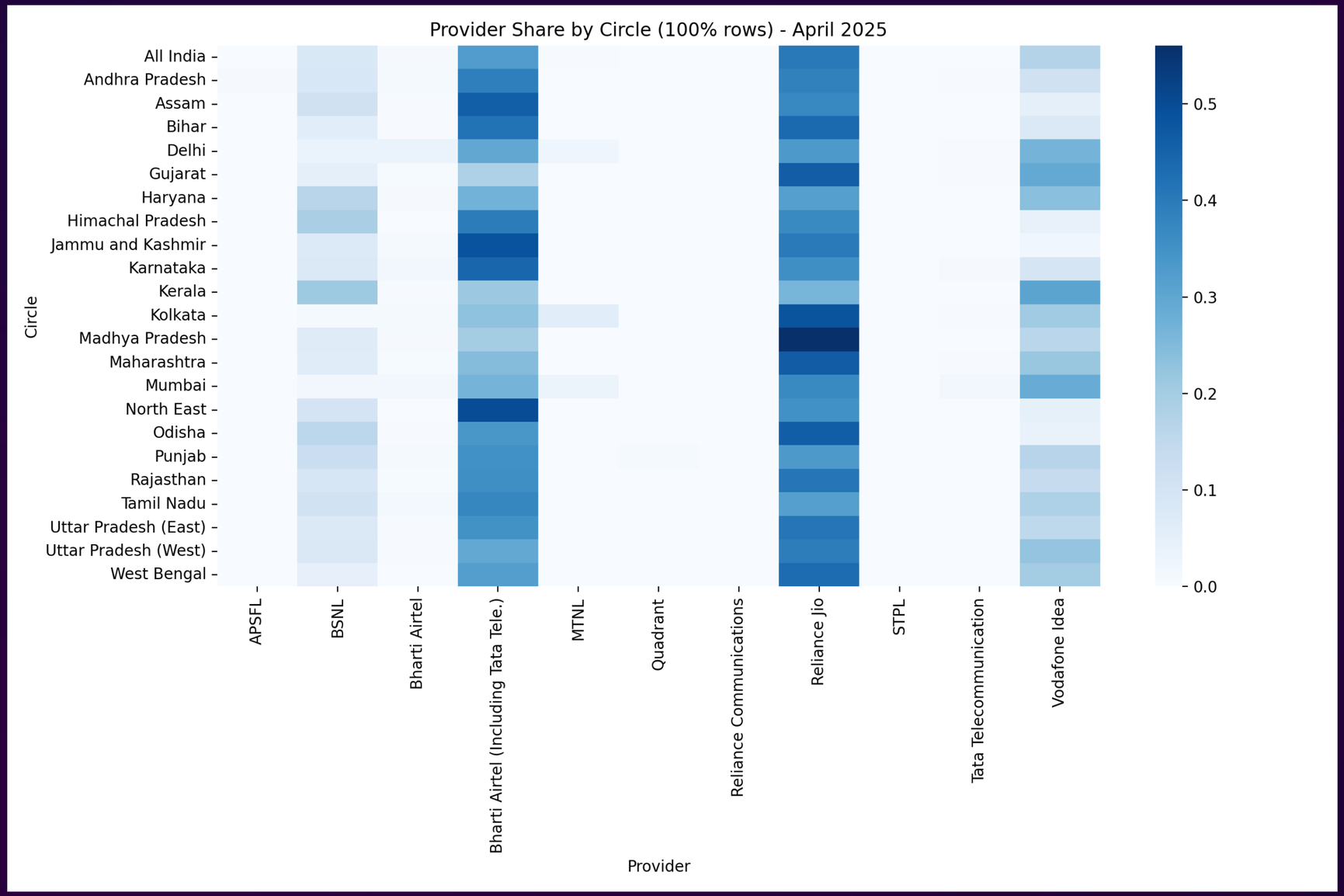




Total Subscribers	Average Subscribers	Total circles	Total Providers
407bn	5.75M	28	42



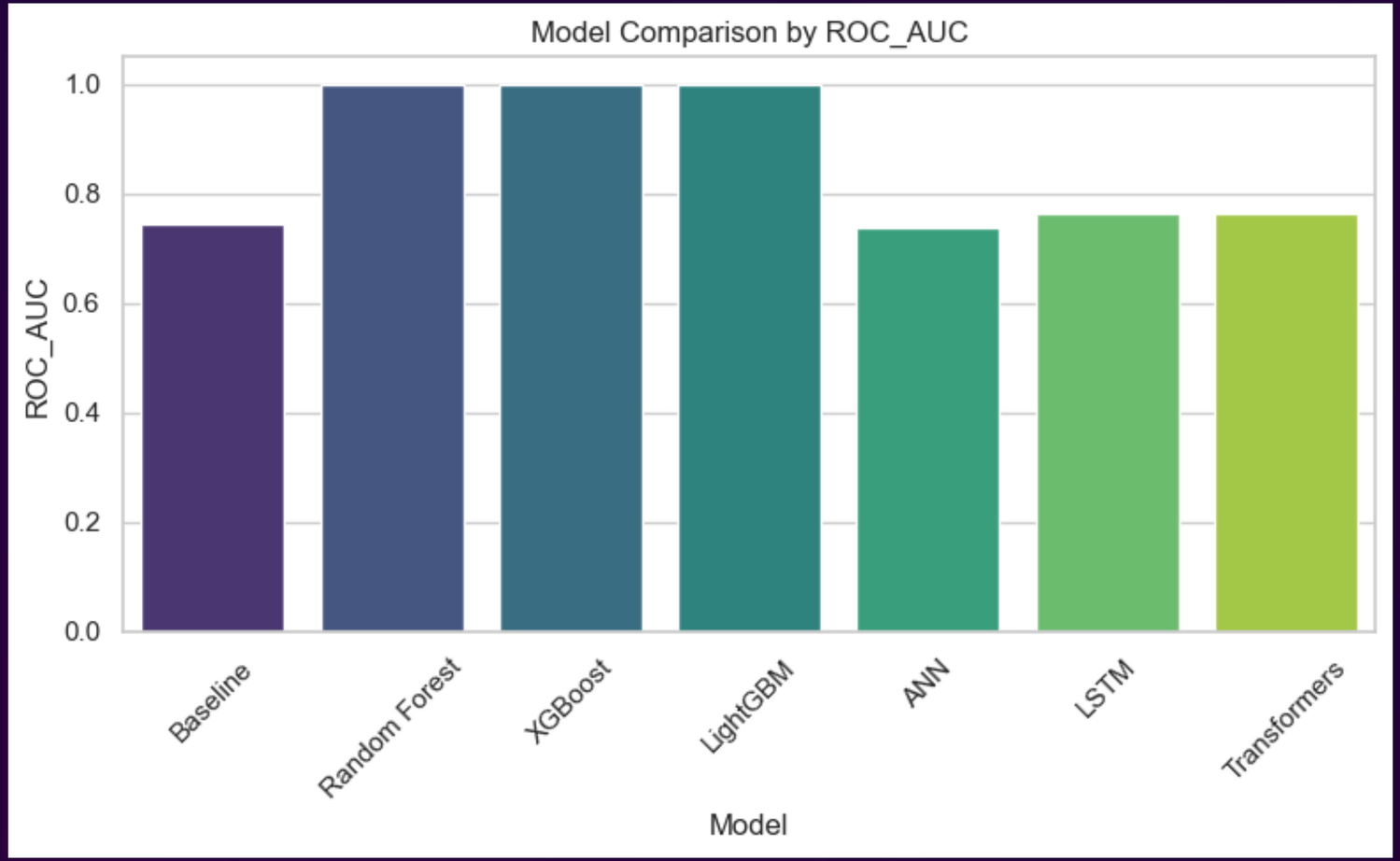



Provider Share by Circle (100% rows) - April 2025
Circle
All India
Andhra Pradesh
Assam
Bihar
Delhi
Gujarat
Haryana
Himachal Pradesh
Jammu and Kashmir
Karnataka
Kerala
Kolkata
Madhya Pradesh
Maharashtra
Mumbai
North East
Odisha
Punjab
Rajasthan
Tamil Nadu
Uttar Pradesh (East)
Uttar Pradesh (West)
West Bengal
APSFL
BSNL
Bharti Airtel
Bharti Airtel (Including Tata Tele.)
MTNL
Quadrant
Reliance Communications
Reliance Jio
STPL
Tata Telecommunication
Vodafone Idea
Provider
0.0
0.1
0.2
0.3
0.4
0.5

Model Building & Training

- Built and compared ML models – XGBoost, LightGBM, Random Forest, Logistic Regression.
- Developed DL models – Neural Networks (TensorFlow), LSTM, and Transformers.
- Used Optuna for hyperparameter tuning to boost model accuracy.
- Evaluated performance using Accuracy, Precision, Recall, F1-score, and Confusion Matrix.
- Saved optimized models with Joblib and deployed via Streamlit dashboard.

Model Comparison



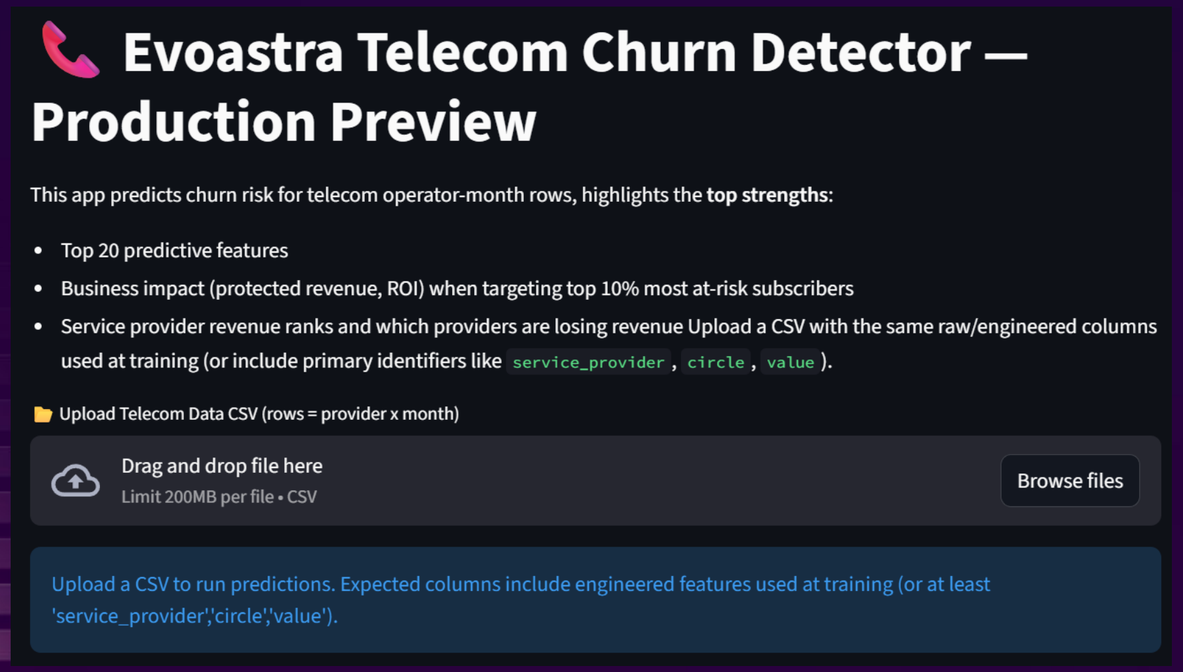
Test Cases:

Introduction :

To ensure the accuracy and robustness of our Proposed System, we designed and conducted multiple test cases. These tests were aimed at validating each stage of the machine learning pipeline, from data input and preprocessing to model prediction and output display in the Streamlit interface.

Test Case 1 : CSV Upload Validation

Input



Screenshot for File Upload

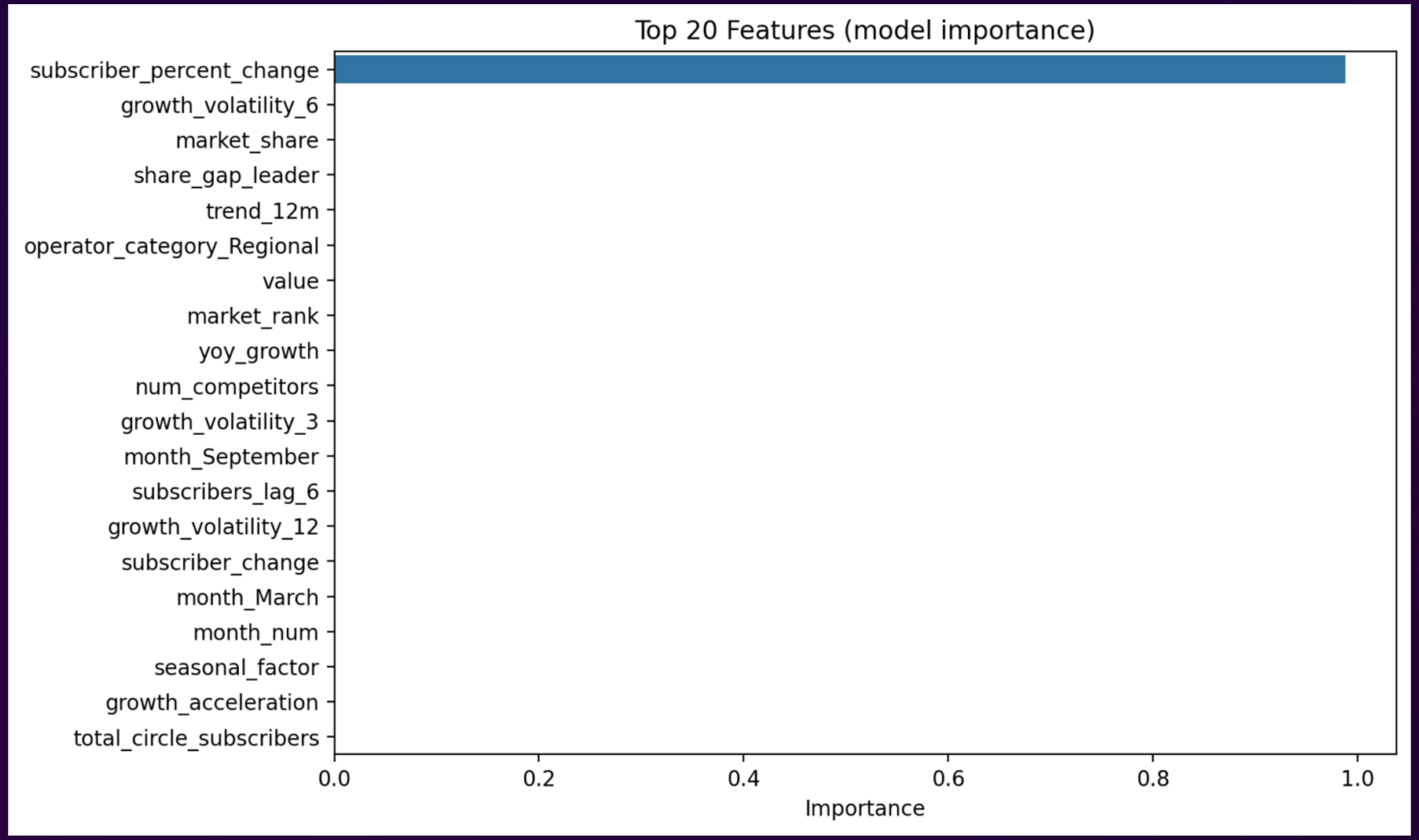
Output

Screenshot after uploading Dataset

Test Case 2 : Dataset Overview

	year	month	circle	type_of_connection	service_provider	value
0	2025	April	Andhra Pradesh	wireless	Bharti Airtel (Including Tata Tele.)	33965795
1	2025	April	Assam	wireless	Bharti Airtel (Including Tata Tele.)	12314102
2	2025	April	Bihar	wireless	Bharti Airtel (Including Tata Tele.)	40967773
3	2025	April	Delhi	wireless	Bharti Airtel (Including Tata Tele.)	18877637
4	2025	April	Gujarat	wireless	Bharti Airtel (Including Tata Tele.)	12401101

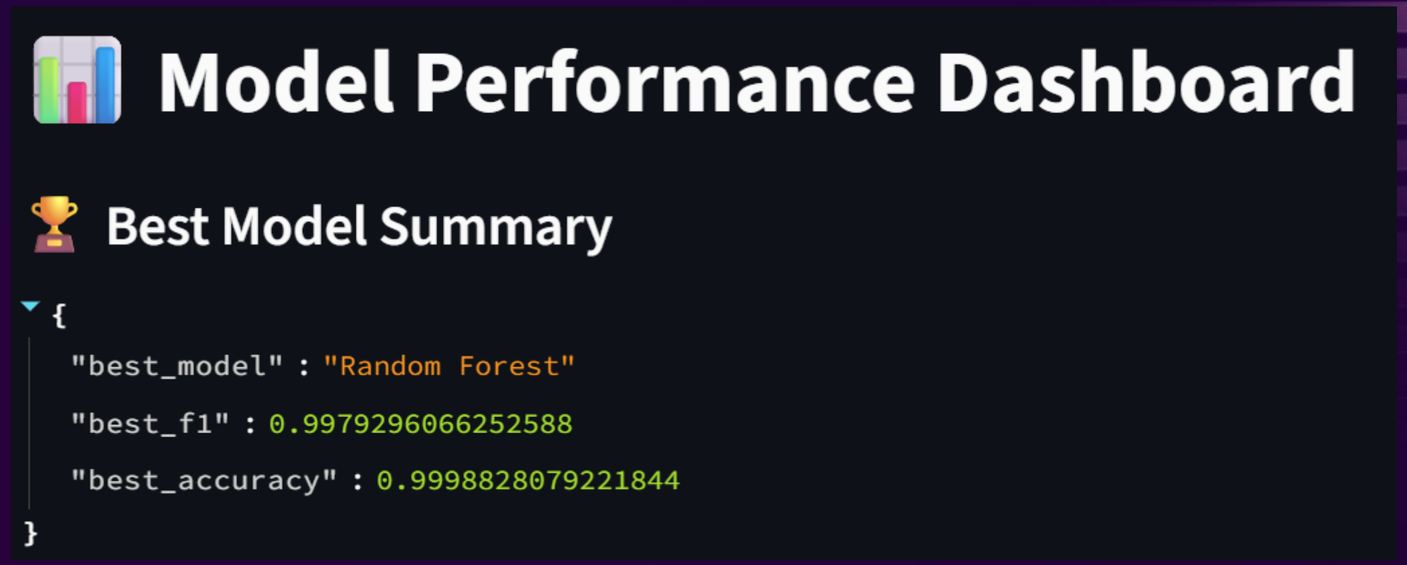
Test Case 3 : Top 20 Model Features



Test Case 4 : Top 20 Model Features (what drives churn predictions)

	feature	importance
0	subscriber_percent_change	0.9881
1	growth_volatility_6	0.0012
2	market_share	0.0008
3	share_gap_leader	0.0008
4	trend_12m	0.0008
5	operator_category_Regional	0.0006
6	value	0.0006
7	market_rank	0.0006
8	yoy_growth	0.0005
9	num_competitors	0.0005
10	growth_volatility_3	0.0005
11	month_September	0.0004
12	subscribers_lag_6	0.0004
13	growth_volatility_12	0.0004
14	subscriber_change	0.0004
15	month_March	0.0004
16	month_num	0.0003
17	seasonal_factor	0.0003
18	growth_acceleration	0.0003
19	total_circle_subscribers	0.0003

Test Case 5 : Best Model Summary



Test Case 6 : Comparison of Models

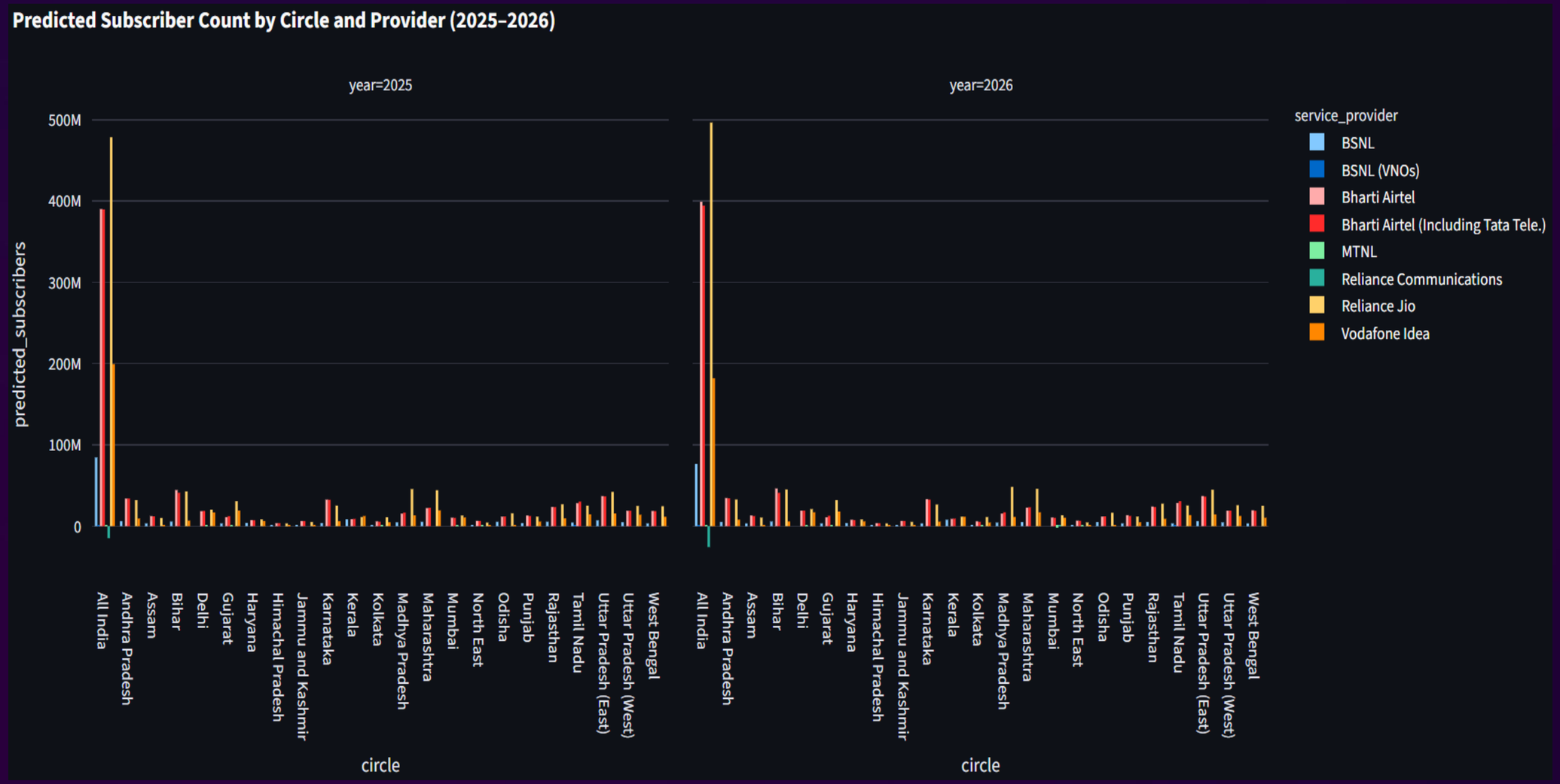
📈 Detailed Comparison of All Models

	accuracy	roc_auc	precision	recall	f1
Baseline	0.860073	0.746198	0.109746	0.556017	0.183311
Random Forest	0.999883	1.000000	0.995868	1.000000	0.997930
XGBoost	0.999531	0.999988	0.983673	1.000000	0.991770
LightGBM	0.998594	0.999963	0.987234	0.962656	0.974790
ANN	0.971054	0.739453	0.483333	0.360996	0.413302
LSTM	0.947381	0.762973	0.229167	0.365145	0.281600
Transformers	0.939646	0.765262	0.213389	0.423237	0.283727

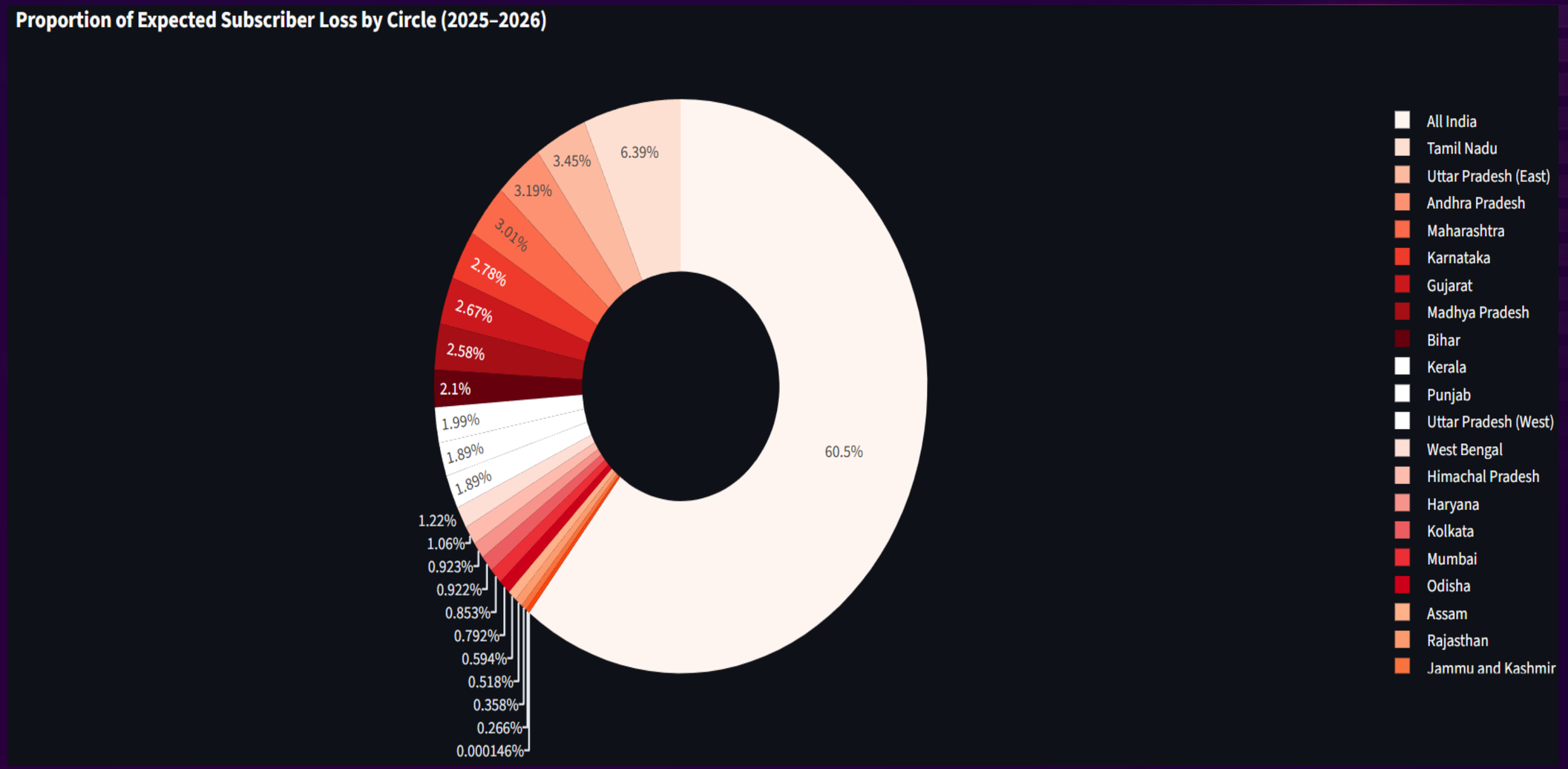
Test Case 7 : Loss Summary (2025–2026)

	service_provider	circle	year	predicted_subscribers	change_in_subscribers	loss_flag
0	BSNL	All India	2025	84428837.3452	-6710005.4048	☑
1	BSNL	All India	2026	76611946.3198	-14526896.4302	☑
2	BSNL	Andhra Pradesh	2025	6033271.1929	-879473.0571	☑
3	BSNL	Andhra Pradesh	2026	5213998.9177	-1698745.3323	☑
4	BSNL	Assam	2025	3000478.9481	29883.9481	☐
5	BSNL	Assam	2026	2933597.1183	-36997.8817	☑
6	BSNL	Bihar	2025	5631117.9718	-188536.0282	☑
7	BSNL	Bihar	2026	5648204.6428	-171449.3572	☑
8	BSNL	Delhi	2025	0	0	☐
9	BSNL	Delhi	2026	0	0	☐
10	BSNL	Gujarat	2025	2746969.362	-538310.388	☑
11	BSNL	Gujarat	2026	1939857.8108	-1345421.9392	☑
12	BSNL	Haryana	2025	3980042.0252	-221470.9748	☑
13	BSNL	Haryana	2026	3734671.6354	-466841.3646	☑
14	BSNL	Himachal Pradesh	2025	1222425.0252	-454956.9748	☑

Test case 8 : Subscriber Trends (2025–2026)



Test Case 9 : Circles Expected Loss (2025–2026)



Challenges Faced

- **Dirty Subscriber Data:** The critical value column had mixed data types and missing entries (≈12,332), requiring force conversion and standardization by replacing non-numeric text with 0.
- **Categorical Inconsistency:** Inconsistent naming across regions and operators (e.g., 50 unique circle names) risked feature explosion, requiring standardization and conversion to the category data type.
- **Time-Series Foundation:** Month-over-Month calculations were blocked by the lack of a single standardized date column, necessitating manual parsing and combination of the separate year and month fields.
- **Overfitting Risk:** Initial cross-validation of powerful models (Random Forest and XGBoost) yielded 1.0000 ROC-AUC scores, demanding rigorous tuning to ensure the final model generalizes.
- **High Dimensionality:** One-Hot Encoding the 42 service_providers and other categories ballooned the input to 113 features, complicating model training and increasing sparsity.
- **Acute Class Imbalance (The "Not Enough Data" Problem):** The target variable was severely skewed (≈80% "No Churn" vs. ≈20% "Churn"), meaning the data wasn't sufficient to show the model what churn looks like, risking a uselessly accurate predictor; this required specialized techniques (like class weighting and F1-Score focus) to overcome.

Useful Insights

- **Financial Impact:** Maximized ROI with a projected net annual benefit of ₹ 5,35,750 and an ROI of 1,256.15% by targeting the top 10% of high-risk customers.
- **Proactive Planning:** Provided actionable intelligence for network planning (where to invest/divest), marketing campaigns, and customer retention strategies.
- **Actionable Intelligence:** Identified specific regions with consistent subscriber loss trends and detected operator-wise performance variations influenced by regional competition.
- **Future Market Risk:** Provided critical 2025-2026 revenue loss forecasts, helping visualize how market dynamics shift over time through model predictions.
- **Targeted Action:** The system allows for pinpointing the Top N Risky Customers for Action, ensuring resources are applied where they yield the highest retention value.

Future Scope

- ✓ Real-Time Data Pipeline: Integrate live telecom data feeds (e.g., streaming network traffic, call failure rates, or service quality metrics) for real-time churn monitoring, shifting the intervention strategy from monthly prediction to immediate action.
- ✓ Deep Learning for Temporal Dynamics: Incorporate advanced time-series deep learning architectures (Transformers, GRU, or BiLSTM) to improve the capture of complex temporal dynamics and non-linear trends in subscriber behavior.
- ✓ Predictive Financial Modeling: Extend the prediction to include key financial metrics like revenue loss and ARPU (Average Revenue Per User), enabling comprehensive quantified risk assessment.
- ✓ Prescriptive Recommendations: Develop a Recommendation Engine that suggests specific, context-aware retention actions (e.g., promotional campaigns) to prevent churn in high-risk regions.
- ✓ Deployment & Integration: Operationalize the system by deploying the model as a web-based API service for seamless integration into the telecom analytics and planning teams.

Conclusion

- This project successfully demonstrates the power of machine learning and data-driven intelligence in predicting market-level telecom churn.

- By precisely leveraging TRAI data, the final Random Forest model (with a **0.9979 F1-Score**) identifies the regions and operators most at risk of subscriber loss. This capability moves the business from reactive to strategic intervention, directly translating predictive accuracy into measurable financial gain:

  - A projected net annual benefit with an impressive 1,256.15% ROI.

  - This reinforces the competitive position by protecting core revenue and market share.

Thank You