

# Telecom Customer Churn Prediction

*A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree*

*of*

**Bachelor of Technology**

**in The Department of AI & DS**

**Big Data Analytics-22DSB3303A**

Submitted by

<b>2210080068</b>	<b>:</b>	<b>V. Rohan</b>
<b>2210080032</b>	<b>:</b>	<b>Gundelly Siddartha Yadav</b>
<b>2210080018</b>	<b>:</b>	<b>S. Shanmukh</b>

Under the guidance of

**Dr. Shahin Fatima**



Department of Artificial intelligence and Data Science

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075

FEB - 2025.

## **Introduction**

In the highly competitive telecommunications industry, customer retention presents a significant challenge. Customer churn, the process through which customers discontinue their services, directly impacts the profitability and sustainability of telecommunications companies. Predicting customer churn is essential as it enables businesses to take proactive measures to retain customers by addressing their concerns and enhancing service quality.

Telecommunications companies generate substantial amounts of customer data, encompassing call records, internet usage, billing details, and customer service interactions. By analyzing this data utilizing machine learning and predictive analytics, telecommunications operators can gain valuable insights into customer behaviour, enabling them to identify at-risk customers and implement timely interventions. The primary objective of this project is to develop a Telecom Customer Churn Predictor employing machine learning techniques. This system will analyse customer data to classify customers as either predisposed to churn or likely to remain, thereby assisting telecommunications companies in optimizing their retention strategies.

## **Literature Review and Related Work**

Customer churn prediction has been an active area of research in data science and machine learning. Various studies have employed statistical methods, machine learning algorithms, and deep learning models to predict churn with high accuracy.

1. Traditional Statistical Methods: Early approaches relied on statistical models like logistic regression and decision trees. These methods provided interpretable results but often lacked the ability to capture complex relationships in large datasets.
2. Machine Learning Approaches: More recent studies have utilized Support Vector Machines (SVM), Random Forest, Gradient Boosting Machines (GBM), and Artificial Neural Networks (ANN) to enhance prediction accuracy.

3. Deep Learning Models: Advanced techniques such as Recurrent Neural Networks (RNNs) and Transformer-based architectures like BERT have been explored for analyzing customer behaviour and sentiment from text-based interactions.
4. Hybrid Models: Some research has focused on combining multiple models to improve churn prediction performance, leveraging ensemble learning techniques.

## **Methodology**

The proposed churn prediction system will follow a structured data analysis pipeline consisting of data collection, preprocessing, feature engineering, model training, and evaluation.

### **1. Data Collection**

The dataset for this project will include various customer attributes such as:

- Demographic Data: Age, gender, location
- Service Usage Data: Call duration, internet usage, roaming frequency
- Billing Information: Monthly charges, total charges, payment method
- Customer Support Interaction: Number of complaints, resolution time
- Contract Details: Subscription type, contract length, tenure

### **2. Data Preprocessing**

Data preprocessing is essential for ensuring data quality and preparing it for model training. The following steps will be performed:

- Handling missing values through imputation techniques.
- Encoding categorical variables using One-Hot Encoding or Label Encoding.
- Scaling numerical features using Min-Max Scaling or Standardization.
- Removing outliers using statistical techniques like the Interquartile Range (IQR) method.

### **3. Feature Engineering**

Feature engineering plays a crucial role in improving model performance. Key techniques include:

- Creating new features such as average call duration per month.

- Aggregating service usage data over different time periods.
- Identifying customer engagement patterns based on past interactions.

#### 4. Model Selection and Training

Several machine learning models will be evaluated to determine the best-performing algorithm for churn prediction:

- Logistic Regression: A simple yet effective baseline model.
- Decision Trees and Random Forests: Useful for capturing non-linear relationships.
- Gradient Boosting Algorithms (XGBoost, LightGBM, CatBoost): Highly effective for structured data.
- Deep Learning (Artificial Neural Networks): For capturing complex patterns in high-dimensional data.

Each model will be trained and validated using techniques such as cross-validation and hyperparameter tuning to optimize performance.

### **Model Evaluation**

Model performance will be assessed using various evaluation metrics:

- Accuracy: Measures the overall correctness of predictions.
- Precision and Recall: Evaluates the model's ability to correctly identify churners.
- F1-Score: Balances precision and recall for better performance assessment.
- AUC-ROC Curve: Measures the model's ability to distinguish between churners and non-churners.
- Applications of Telecom Churn Prediction

Churn prediction has several real-world applications in the telecom industry, including:

#### 1. Customer Retention Strategies:

- Proactively offering personalized discounts and promotions to high-risk customers.
- Enhancing customer service responsiveness for at-risk segments.

## 2. Revenue Optimization:

- Reducing customer acquisition costs by retaining existing customers.
- Identifying high-value customers and ensuring their satisfaction.

## 3. Service Quality Improvement:

- Using churn insights to enhance network performance and reliability.
- Addressing common customer complaints before they lead to churn.

## 4. Marketing Campaign Optimization:

- Targeting customers with retention-focused marketing messages.
- Designing customized offers based on churn probability scores.

## **Challenges in Churn Prediction**

While predictive analytics can provide valuable insights, several challenges must be addressed:

### 1. Imbalanced Datasets:

- Churners often represent a small fraction of the customer base, leading to class imbalance issues. Techniques like oversampling, under sampling, or Synthetic Minority Over-sampling Technique (SMOTE) can help mitigate this.

### 2. Data Privacy and Compliance:

- Handling sensitive customer data requires adherence to privacy regulations such as GDPR and CCPA.

### 3. Evolving Customer Behaviours:

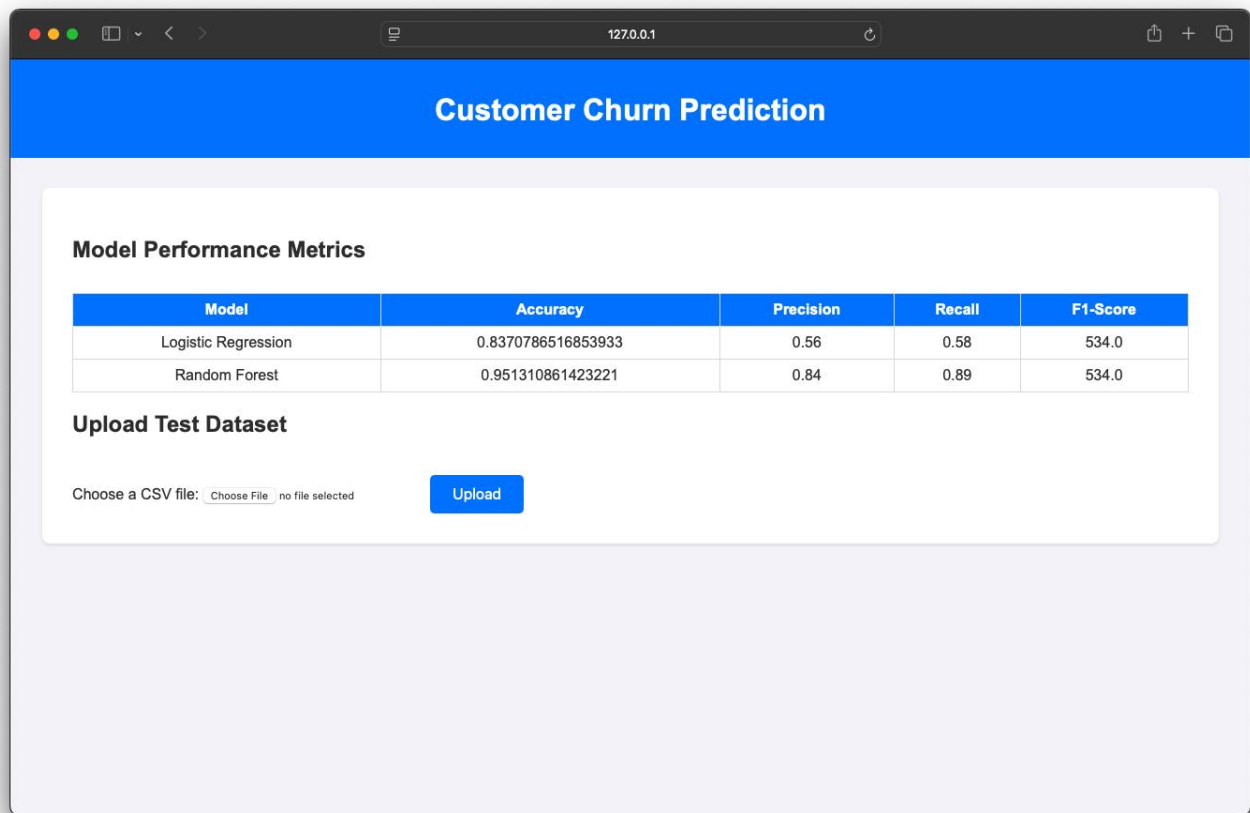
- Customer preferences change over time, requiring models to be periodically retrained with updated data.

### 4. Feature Selection Complexity:

- Selecting the right set of features is crucial for building an effective model without introducing noise or overfitting.

## Application Development

Till date we have collaborated and made a flask based application where the predictions are done when given a test dataset. Examples of the site below:



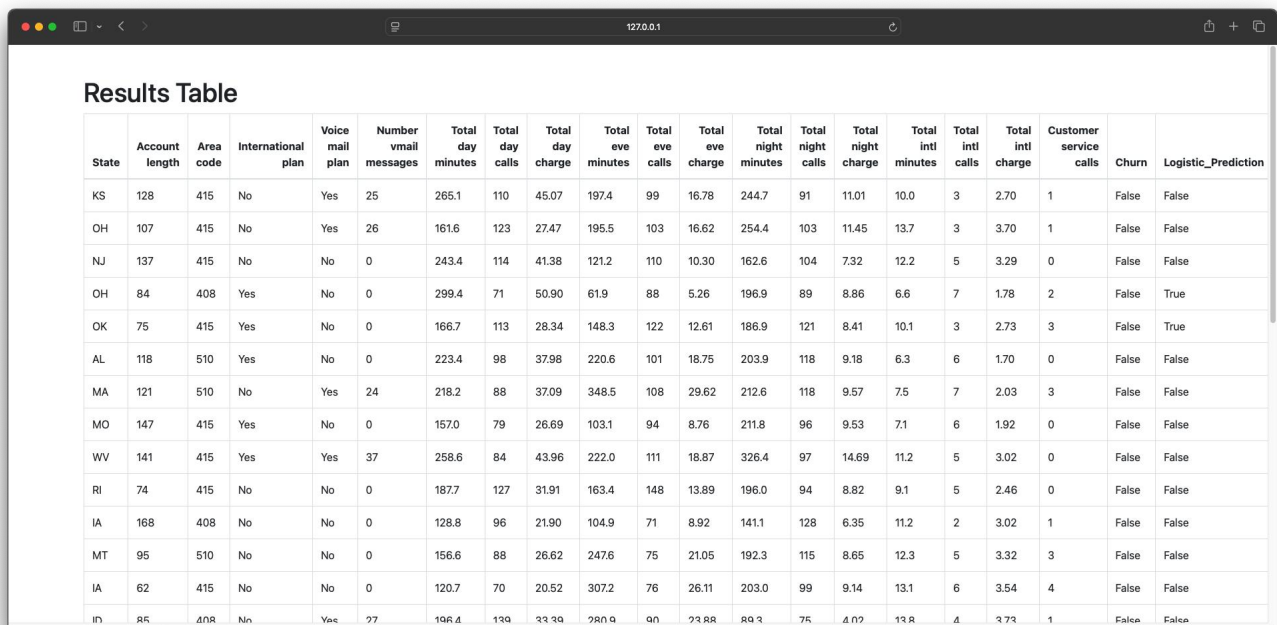
The screenshot displays a web application titled "Customer Churn Prediction". It features a blue header bar with the title. Below the header, there is a section titled "Model Performance Metrics" containing a table with five columns: Model, Accuracy, Precision, Recall, and F1-Score. The table lists two models: Logistic Regression and Random Forest. Below this table is a section titled "Upload Test Dataset" which includes a file upload interface with a "Choose a CSV file:" label, a "Choose File" button, and an "Upload" button.

### Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.8370786516853933	0.56	0.58	534.0
Random Forest	0.951310861423221	0.84	0.89	534.0

### Upload Test Dataset

Choose a CSV file:  no file selected



The screenshot displays a web application titled "Results Table". It features a table with 21 columns: State, Account length, Area code, International plan, Voice mail plan, Number vmail messages, Total day minutes, Total day calls, Total day charge, Total eve minutes, Total eve calls, Total eve charge, Total night minutes, Total night calls, Total night charge, Total intl minutes, Total intl calls, Total intl charge, Customer service calls, Churn, and Logistic\_Prediction. The table contains 15 rows of data, each representing a customer record with various attributes and their predicted churn status.

### Results Table

State	Account length	Area code	International plan	Voice mail plan	Number vmail messages	Total day minutes	Total day calls	Total day charge	Total eve minutes	Total eve calls	Total eve charge	Total night minutes	Total night calls	Total night charge	Total intl minutes	Total intl calls	Total intl charge	Customer service calls	Churn	Logistic_Prediction
KS	128	415	No	Yes	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10.0	3	2.70	1	False	False
OH	107	415	No	Yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.70	1	False	False
NJ	137	415	No	No	0	243.4	114	41.38	121.2	110	10.30	162.6	104	7.32	12.2	5	3.29	0	False	False
OH	84	408	Yes	No	0	299.4	71	50.90	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2	False	True
OK	75	415	Yes	No	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3	False	True
AL	118	510	Yes	No	0	223.4	98	37.98	220.6	101	18.75	203.9	118	9.18	6.3	6	1.70	0	False	False
MA	121	510	No	Yes	24	218.2	88	37.09	348.5	108	29.62	212.6	118	9.57	7.5	7	2.03	3	False	False
MO	147	415	Yes	No	0	157.0	79	26.69	103.1	94	8.76	211.8	96	9.53	7.1	6	1.92	0	False	False
WV	141	415	Yes	Yes	37	258.6	84	43.96	222.0	111	18.87	326.4	97	14.69	11.2	5	3.02	0	False	False
RI	74	415	No	No	0	187.7	127	31.91	163.4	148	13.89	196.0	94	8.82	9.1	5	2.46	0	False	False
IA	168	408	No	No	0	128.8	96	21.90	104.9	71	8.92	141.1	128	6.35	11.2	2	3.02	1	False	False
MT	95	510	No	No	0	156.6	88	26.62	247.6	75	21.05	192.3	115	8.65	12.3	5	3.32	3	False	False
IA	62	415	No	No	0	120.7	70	20.52	307.2	76	26.11	203.0	99	9.14	13.1	6	3.54	4	False	False
IN	85	408	No	Yes	27	196.4	130	33.30	280.0	90	23.88	89.3	75	4.02	13.8	4	3.73	1	False	False

## Future Directions

Future work in telecom churn prediction can explore the following areas:

- Real-Time Churn Prediction: Implementing streaming analytics to provide real-time churn alerts.
- Explainable AI (XAI): Developing interpretable machine learning models to explain churn decisions.
- Sentiment Analysis Integration: Incorporating customer feedback and sentiment data from social media and customer support interactions.
- Reinforcement Learning for Retention Strategies: Using reinforcement learning to dynamically adjust customer retention strategies based on model predictions.

## References:

1. F. Provost and T. Fawcett, *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*, O'Reilly Media, 2013.
2. C. C. Aggarwal, *Data Classification: Algorithms and Applications*, CRC Press, 2014.
3. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.
4. G. E. Batista, A. L. C. Bazzan, and M. C. Monard, "Balancing Training Data for Automated Annotation of Keywords: A Case Study," in *Proceedings of the Brazilian Symposium on Artificial Intelligence*, 2003, pp. 10-18.
5. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology Behind Search*, Addison-Wesley, 2011.
6. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
7. H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, 2009.
8. J. A. Berry and G. S. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, John Wiley & Sons, 2011.
9. Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436-444, 2015.

10.C. Elkan, "The Foundations of Cost-Sensitive Learning," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2001, pp. 973-978.