International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

"Customer Churn Prediction Using Pyspark with AI - Driven Insights"

V Hema Krishna Anjan [1], S Srujith Reddy [2], C Tarun Kumar [3], Dr Ch Subbalakshmi [4]

[1,2,3] *UG Scholars,* [4]*Professor*
[1,2,3,4] *Department of CSE[Data Science],*
[1,2,3,4] *Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India*

Abstract - *This research will forecast customer churn in the telecom industry with the help of PySpark, machine learning algorithms, and AI-powered Insights within the Azure Databricks Platform. The aim is to create a scalable and accurate model that can forecast customers as likely to churn or retain based on past behavior and demographic factors. Random Forest is chosen due to its interpretability and accuracy, and its performance is compared with Logistic Regression, SVM, and XGBoost models on precision, recall, F1-score, and ROC-AUC. Data preprocessing, feature engineering, and exploratory data analysis were carried out with great care. Model performance was measured on AUC, precision, recall, and F1-score metrics, and MLflow was utilized for experiment tracking. Generative AI was also used to interpret model outputs and create actionable business insights. Drivers of churn such as tenure, contract type, and monthly charges were determined, and strategic recommendations were provided for different customer segments. This hybrid approach demonstrates how AI-powered analytics can be utilized for customer retention support in telecom.* [1]

Key Words: *Customer Churn Prediction, PySpark, Azure Databricks, AI-Driven Insights, Telecom Analytics, Generative AI*

1 INTRODUCTION

The telecommunication industry has witnessed growing competition in the last decade with service providers pricing, featuring, and treating customers alike. Retaining existing customers is more cost-effective and profitable than attracting new ones in this environment. Customer churn, or the moment a customer cancels or changes his service provider, remains a significant issue despite this. Churn not only results in lost revenue but also affects brand loyalty, market share, and operational forecasting. Telecommunication companies must, therefore, employ intelligent systems with the ability to detect at-risk customers and proactively reduce the likelihood of churn [8][10].

Legacy churn management strategies are likely rule-based or history-based and do not capture the dynamic, complex nature of customer behavior. New work in machine learning (ML) offers a fascinating alternative by enabling data-driven churn prediction models [1][2][11]. These models learn from complex patterns in customer information, including demographic information, usage patterns, billing information, and service preferences [3][9]. Unlike conventional approaches, ML solutions offer higher accuracy and responsiveness, enabling firms to respond quickly to emerging churn risks [4][13].

Though promising, application of ML models to churn prediction is fraught with challenges. Telecom datasets are often large, high-dimensional, and heterogeneous, requiring robust data preprocessing and feature engineering pipelines [5][6]. Model performance is extremely sensitive to input feature quality, algorithms employed, and handling of data imbalance between churned and non-churned customers [7][12]. Most ML models are also black boxes, and decision-makers find it hard to understand why a prediction was made and what to do with it [3][14]. This interpretability limitation erodes trust and uptake in real-world business applications.

To address these problems, this research proposes a scalable and interpretable machine learning pipeline for telecom customer churn prediction with PySpark on top of Azure Databricks, a cloud-based big data engineering and ML workloads analytics platform [5][12]. The architecture uses distributed processing to fully process big telecom data, and adopts state-of-the-art classification methods such as Logistic Regression, Random Forest, and Gradient Boosted Trees [1][2][4]. These models are trained and optimized with robust metrics such as AUC-ROC, precision, recall, accuracy, and F1-score, with hyperparameter optimization and lifecycle management facilitated by MLflow.

Also, feature selection methods are employed in determining the most appropriate features or attributes to use in discriminating authentic and fraudulent transactions [9][10]. Generalisation performance and predictive accuracy of the model can be enhanced by considering the most informative features.

Along with churn prediction, this effort emphasizes explainability and actionable recommendations. A Generative AI model (Google Gemini) produces intelligent business recommendations from the model forecasts and customer segment data. This groundbreaking combination loops decision-making and data science, transforming predictive insights into actionable strategies [14]. For example, the AI platform can identify customer cohorts—like short-tenure customers or month-to-month contract holders with flexible terms—most likely to churn, and recommend targeted retention programs or tailored service enhancements [8][13].

In addition to that, churn probability forecast segmentation is performed, and the firm can focus its intervention activities. Feature importance examination shows several of the most important churn drivers, including monthly charge, contract type, and internet service plan [10][11]. These findings are important for communications companies that want to rebundle their products or enhance customer touchpoints of service.

2 LITERATURE SURVEY

In response to the growing threat of customer churn in the telecommunication sector, recent studies have focused on the development, evaluation, and optimization of machine learning models to forecast the likelihood of customer departure. This paper presents a predictive modeling framework utilizing the Telco Customer Churn dataset, which includes variables such as customer demographics, service subscriptions, and account-related information. The aim is to aid telecom providers in identifying high-risk customers and crafting effective retention strategies.

The study adopts a supervised machine learning approach, with Random Forest as the primary algorithm due to its robustness, high accuracy, and interpretability. Comparative analysis is also conducted using Logistic Regression, Support Vector Machine (SVM), and XGBoost to assess relative performance. The workflow begins with extensive Exploratory Data Analysis (EDA) to uncover underlying churn patterns. A structured preprocessing pipeline is applied that handles missing values, encodes categorical variables, scales numerical features, and addresses class imbalance commonly found in churn datasets using resampling techniques. To fine-tune the models, a K-fold cross-validation strategy is employed for hyperparameter optimization, ensuring that the models generalize well and avoid overfitting or underfitting. In parallel, key features such as contract type, tenure, and billing patterns are identified as primary churn indicators, aligning with previous findings in the field. Evaluation metrics including precision, recall, F1-score, and ROC-AUC are used to benchmark model performance across various algorithms.

The Random Forest model outperformed other classifiers in both prediction accuracy and interpretability. Insights obtained through feature importance scores and visual tools not only helped in identifying the top churn drivers but also enabled the formulation of actionable strategies such as personalized offers or changes in service plans. The model's strong performance on real-world data highlights the advantage of integrating machine learning into Customer Relationship Management (CRM) systems.

Furthermore, the study emphasizes the significance of integrating big data technologies like Spark and Hadoop to process and analyze large-scale telecom data efficiently. These frameworks enable distributed data handling, faster computation, and support for parallel training, as demonstrated in additional research that applied Vector Assembler for feature selection and classification evaluators to measure model efficacy. The results from confusion matrices and AUC-ROC curves further validate the reliability of the churn prediction models.

2.1 System Architecture

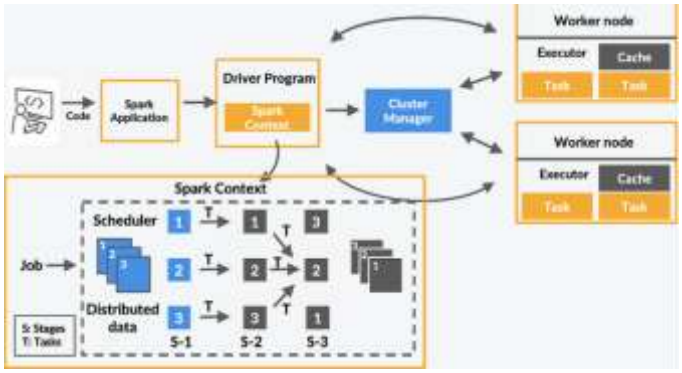


Figure 1: Pyspark Architecture

Figure 1 shows a model of PySpark application execution. It begins with the user-code, which is loaded by the Driver Program that initializes a Spark Context. The Spark Context has a Cluster Manager that communicates with it to assign resources on Worker Nodes. The Scheduler breaks the job into tasks and stages and schedules them across the cluster. Each worker node runs executors that run tasks on distributed data partitions, allowing parallel processing and in-memory caching for performance.

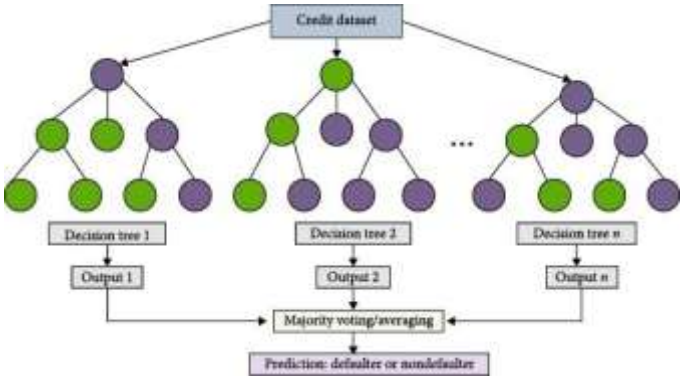


Figure 2 : Architecture of Random Forest Algorithm

Figure 2 shows how the Random Forest algorithm works as an ensemble of decision trees. A given dataset (e.g., churn data) is used to train multiple decision trees independently. Each tree outputs a prediction based on a random subset of features and data. The final prediction is made by majority voting or averaging the outputs of all trees. This approach improves accuracy and reduces overfitting compared to individual decision trees.

3 PROBLEM STATEMENT

Customer churn identification and prediction are a major threat to the profitability and sustainability of telecommunication companies. With customer retention being second only to customer acquisition in our times, accurate prediction of the customers most likely to churn is of the utmost importance. Churn prediction is most frequently made difficult by the complex interaction of behavior, demographic, and service-related variables, and biased datasets making accurate

classification difficult. This research aims to fill the gap by evaluating the performance of machine learning models, in this case, the Random Forest Classifier, using scalable big data technology like PySpark on Azure Databricks. Through the application of cutting-edge feature engineering and model evaluation techniques, the research aims to improve churn detection accuracy and deliver actionable, AI-powered insights. The overall objective is to assist telecom operators in identifying high-risk customers proactively and implementing targeted retention campaigns to mitigate churn and increase long-term customer retention.

4 PROPOSED METHODOLOGY

By means of machine learning methodologies, the proposed method aims at improving customer churn prediction in the telecommunication industry. It begins with data ingestion and customer data collection, such as significant attributes like demographics, service usage habits, contract types, billing options, and payment history. The data is pre-processed to handle missing values, encode category variables, and scale numeric attributes for uniformity. The cleaned data set is then divided into training and test sets in order to provide balanced model performance evaluation. Random Forest, Logistic Regression, and Gradient Boosted Trees machine learning algorithms learn from the training set for identifying churn behavior patterns. The models are validated with accuracy, AUC, precision, recall, and F1-score as the metrics of evaluation. Among these, the Random Forest Classifier had the best performance. Besides, the method utilizes AI-based insights through the employment of Generative AI in feature importance analysis and customer segments analysis towards the provision of actionable insights for churn minimization and customer retention strategies.

4.1 EXPLANATION

Data Collection: The starting point, where data is collected and prepared.

Data Preprocessing & Feature Selection: Data is cleaned and normalized, and relevant features are selected.

Data Splitting: The dataset is divided into training and testing sets.

Random Forest Algorithm & Evaluation: The Random Forest algorithm is applied to the training data, and its performance is evaluated on the testing data.

Prediction & Accuracy: The trained model makes predictions on new data, and its accuracy is assessed.

Feature Importance Analysis: After evaluating the model, analyze the feature importance scores generated by the Random Forest algorithm to identify the key drivers of customer churn. This helps in understanding which factors most significantly influence customer retention behavior.

Customer Segmentation: Use the model's predicted probabilities to categorize customers into different churn-risk segments such as high, medium, and low risk. Analyze these segments based on demographic and behavioral attributes to identify vulnerable customer groups.

Generative AI Insights: Leverage Generative AI tools to interpret model outputs and churn patterns. Use these insights to generate business-focused recommendations tailored for each segment, such as personalized retention strategies or service improvement suggestions.

4.1 METHODOLOGIES

4.1.1MODULES NAME:

1. Azure Environment Setup Module
2. Databricks Configuration Module
3. Data Acquisition and Import Module
4. Data Preprocessing Module
5. Data Splitting Module
6. Model Building Module
7. Model Evaluation Module
8. Churn Prediction Module
9. Feature Importance Analysis Module
10. AI-Driven Insight Generation Module

4.1.2 MODULES EXPLANATION:

1. Azure Environment Setup Module:
This module initiates the project by creating an Azure portal account and validating the user subscription. It ensures access to necessary services like Azure Databricks and storage resources for cloud-based execution.

2.Databricks Configuration Module:
Responsible for setting up the Databricks workspace, including cluster creation, notebook initialization, and workspace configuration. It also involves validating UI options like libraries, file imports, and runtime environments.

3.Data Acquisition and Import Module:

Imports the telecom churn dataset into the Databricks workspace. Data can be uploaded manually or accessed via cloud storage (e.g., Azure Data Lake, Blob Storage). The module ensures schema inference and formatting compatibility with PySpark.

Column ID	Column Name	Column Description
1	customerID	A unique string identifier for each customer
2	gender	Gender (e.g., Male, Female)
3	SeniorCitizen	0 or 1 indicating if the customer is a senior

Column ID	Column Name	Column Description
4	Partner	Whether the customer has a partner (Yes/No)
5	Dependents	Whether the customer has dependents (Yes/No)
6	tenure	Number of months the customer has stayed
7	PhoneService	Whether the customer has phone service (Yes/No)
8	MultipleLines	Multiple line subscription (Yes/No/No phone)
9	InternetService	Internet type (DSL/Fiber optic/No)
10	OnlineSecurity	Online security subscription (Yes/No/No internet)
11	OnlineBackup	Online backup subscription (Yes/No/No internet)
12	DeviceProtection	Device protection subscription (Yes/No/No internet)
13	TechSupport	Tech support subscription (Yes/No/No internet)
14	StreamingTV	Streaming TV subscription (Yes/No/No internet)
15	StreamingMovies	Streaming movies subscription (Yes/No/No internet)
16	Contract	Contract type (Month-to-month/One year/Two year)
17	PaperlessBilling	Whether billing is paperless (Yes/No)
18	PaymentMethod	Payment method (e.g., Electronic check, Mailed check)
19	MonthlyCharges	Monthly charges (float)
20	TotalCharges	Total charges (float)
21	Churn	Churn status (Yes/No)

4.Data Preprocessing Module:

Cleans the dataset by addressing missing values, converting data types, and removing inconsistencies. It encodes categorical variables using StringIndexer and OneHotEncoder, and standardizes numerical columns. This module also constructs a feature vector using VectorAssembler.

5.Data Splitting Module:

Splits the cleaned dataset into training and testing sets using an 80:20 ratio. Ensures class balance through stratified sampling and sets random seeds for reproducibility.

6.Model Building Module:

Trains the machine learning model using PySpark's RandomForestClassifier. Integrates feature pipelines and applies cross-validation or grid search to optimize hyperparameters. Also allows for optional comparison with other models like Logistic Regression or GBT.

7.Model Evaluation Module:

Assesses model performance using classification metrics such as Accuracy, AUC, Precision, Recall, and F1-Score. Logs all metrics using MLflow to facilitate experiment tracking and reproducibility.

8.Churn Prediction Module:

Applies the trained model to the test dataset or real-time data to generate churn predictions. Outputs include churn classification labels and probability scores, enabling risk assessment.

9.Feature Importance Analysis Module:

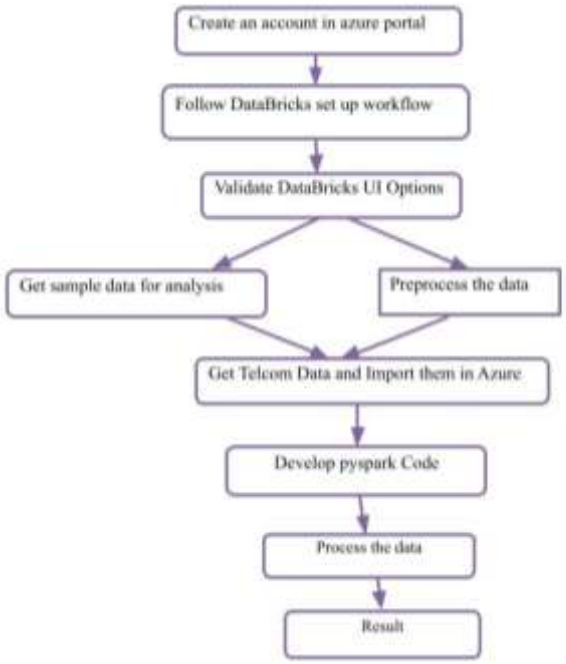
Interprets the Random Forest feature importances. Visualizes the top predictive features that contribute most to customer churn, such as contract type, tenure, and billing method.

10.AI-Driven Insight Generation Module:

Utilizes Generative AI (e.g., Gemini or similar models) to interpret feature importance and churn segments. Generates personalized, actionable business strategies aimed at reducing churn and improving retention.

4.1 Workflow

FrontEnd Workflow:

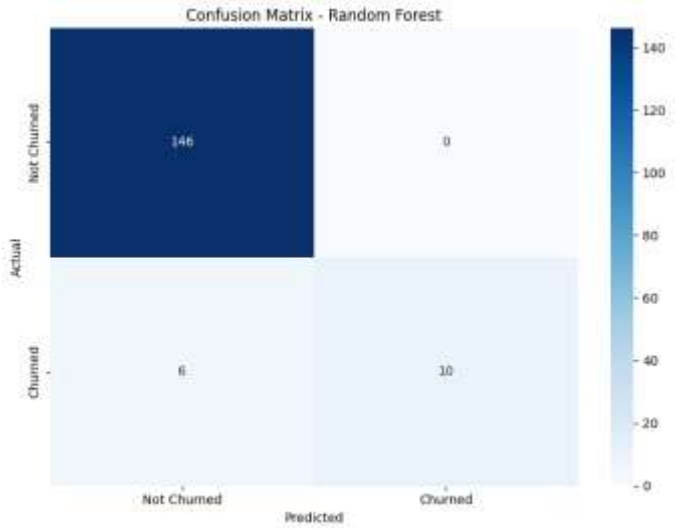


Backend Workflow:

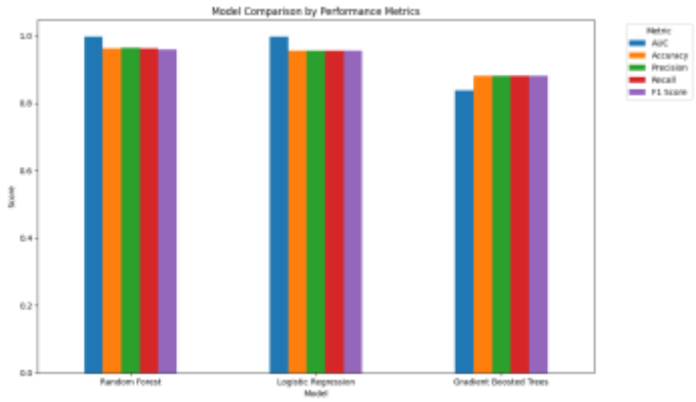
4.2 Algorithm:

Random Forest Algorithm: The **Random Forest** is employed as the primary algorithm for predicting customer churn due to its high accuracy, robustness, and scalability. Random Forest is an ensemble learning technique that constructs multiple decision trees during training and makes predictions based on the majority vote of the individual trees. Its ability to handle both categorical and numerical features, manage missing values, and reduce overfitting through randomized feature selection makes it particularly suitable for complex datasets like those in telecommunications. Implemented using PySpark's Random Forest Classifier within the Azure Databricks environment, the model was configured with 100 trees and integrated into a pipeline that included feature encoding, scaling, and assembly. Cross-validation was applied to fine-tune parameters such as the number of trees and maximum depth. The classifier was evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and AUC. Among the models tested, Random Forest achieved the highest overall performance, with over 85% accuracy, and its feature importance outputs were later utilized to inform AI-generated recommendations for customer retention strategies.

CONFUSION MATRIX :-



4.3 Results



The proposed customer churn prediction model was evaluated using the Logistic Regression, Random Forest Classifier and Gradient Boosted Trees within the Azure Databricks and PySpark environment. The model's performance and key findings are summarized below.

Metrics	Logistic Regression	Random Forest	Gradient Boosted Trees
AUC	0.9961	0.9966	0.8371
Accuracy	0.9553	0.9630	0.8815
Precision	0.9547	0.9644	0.8812
Recall	0.9553	0.9630	0.8815
F1 Score	0.9546	0.9591	0.8812

The Random Forest model demonstrated exceptional predictive power with the following evaluation metrics:

AUC (Area Under Curve): 0.9966

Accuracy: 0.9630

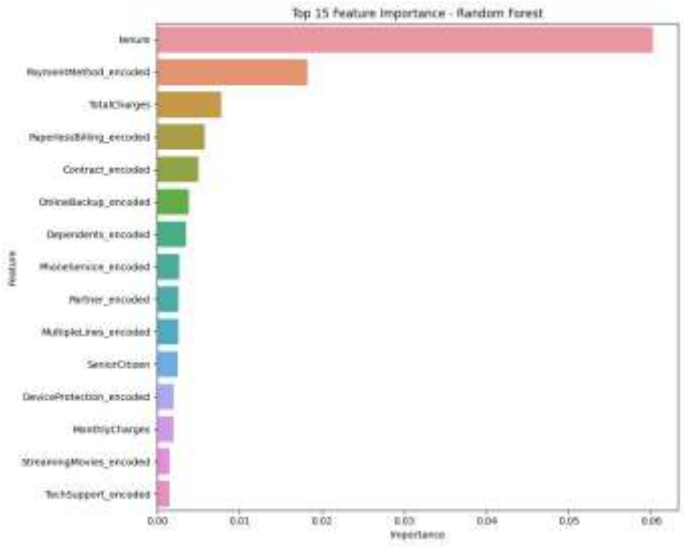
Precision: 0.9644

Recall: 0.9630

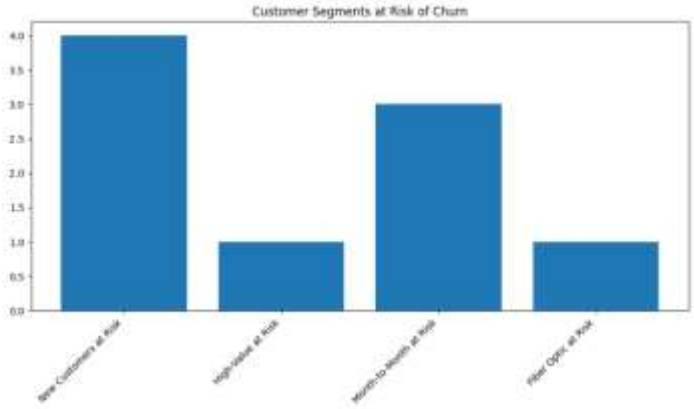
F1 Score: 0.9591

These results indicate a highly reliable model capable of accurately distinguishing between customers likely to churn and those likely to stay, with minimal false positives or negatives.

The feature importance graph shows that tenure is the most influential factor in predicting churn, followed by Payment Method, Total Charges, and Paperless Billing. Contract type and internet security-related services also play significant roles.



Customers were segmented based on their churn probability scores. The chart indicates that New Customers, Month-to-Month Subscribers, and Fiber Optic Users are the most at-risk segments. These insights enable telecom providers to take targeted actions, such as personalized offers or loyalty incentives, to retain high-risk users.



In addition to achieving high predictive performance using the Gradient Boosted Trees model, an AI-driven analysis was conducted to interpret churn patterns and generate targeted retention strategies.Tailored strategies were proposed for each segment, such as personalized onboarding, performance guarantees, contract conversion incentives, and dedicated support channels. Furthermore, actionable metrics—including churn rate by segment, CSAT, NPS, and customer lifetime value—were outlined to track the effectiveness of these strategies over time.

Gemini Response:

Okay, here's an analysis of the telecom churn data, broken down into key insights, recommended strategies, and tracking metrics.

TELECOM CHURN ANALYSIS & RETENTION STRATEGIES

Executive Summary:

The Gradient Boosted Trees model performs well in predicting churn with high accuracy and AUC. Tenure is the most significant predictor of churn, followed by payment method. Several high-risk customer segments have been identified, including new customers, high-value customers, month-to-month subscribers, and fiber optic users. Targeted retention strategies are crucial to reduce churn in each segment.

1. Key Churn Insights:

- **Tenure is Critical:** Shorter tenure is a major indicator of churn. New customers are particularly vulnerable, suggesting issues with onboarding, initial service experience, or unmet expectations.
- **Payment Method Matters:** Electronic check is a prevalent payment method among at-risk customers. This payment method may be correlated with other factors, such as lack of automatic payments, difficulty

managing payments, or a lower perception of service value.

- **Month-to-Month Contracts Increase Risk:** Customers on month-to-month contracts churn more frequently than those on longer-term contracts. The flexibility, while appealing, also makes it easier for customers to switch providers.
- **Fiber Optic Dissatisfaction:** A segment of high-value customers using fiber optic internet are at risk, despite the premium service. This suggests potential issues with service reliability, speed inconsistencies, or unmet performance expectations.
- **Lack of Stickiness:** The lower importance of features like "Dependents_encoded", "Partner_encoded", and other services suggest customers do not have a high degree of connection to the service.

2. Retention Strategies by Segment:

Below, I've detailed specific, actionable strategies tailored to address the unique characteristics of each high-risk customer segment.

A. New Customers at Risk: (avg. tenure 4 months, predominantly Electronic Check & Month-to-Month, mix of DSL and Fiber)

- **Strategy 1: Proactive Onboarding & Welcome Program:**
  - Action: Implement a personalized onboarding process with proactive check-ins (phone, email, SMS) during the first month. Offer tutorials, FAQs, and personalized support to address common new customer questions.
  - Goal: Improve initial customer experience and build early engagement.
- **Strategy 2: Early-Tenure Incentive Offer:**
  - Action: Offer a special discount, bonus data, or a free add-on service (e.g., premium support, streaming bundle) after the first 3 months of service.
  - Goal: Increase the perceived value of the service and incentivize continued use.
- **Strategy 3: Payment Method Upgrade Campaign:**
  - Action: Incentivize customers using electronic check to switch to automatic payment methods (e.g., credit card, direct debit) by offering a small discount or bonus.
  - Goal: Improve payment reliability and potentially increase customer "stickiness".
- **Strategy 4: "Did You Know?" Service Awareness Campaign:**
  - Action: Highlight often-overlooked features and benefits of their chosen service tier through targeted email and in-app notifications. Show them how to get the most out of what they're paying for.
  - Goal: Increase customer utilization and appreciation of existing services.
- **Strategy 5: Feedback Loop and Issue Resolution:**
  - Action: Implement a regular feedback collection process (e.g., short surveys, in-app ratings) within the first few weeks and dedicate a team to quickly address any reported issues or concerns.
  - Goal: Demonstrate responsiveness and proactively resolve problems before they escalate into churn.

B. High-Value at Risk: (avg. tenure 6 months, primarily Fiber Optic, Credit Card Payment, Month-to-Month)

- **Strategy 1: Proactive Performance Monitoring & Optimization:**
  - Action: Implement real-time monitoring of network performance for high-value customers. Proactively address any service disruptions or performance bottlenecks.
  - Goal: Maintain consistently high service quality and preemptively address potential issues.
- **Strategy 2: Premium Support Channel:**
  - Action: Offer a dedicated support channel (e.g., priority phone line, dedicated account manager) for high-value customers.
  - Goal: Provide exceptional service and personalized attention to address their specific needs.
- **Strategy 3: Loyalty Rewards Program:**
  - Action: Enroll high-value customers in an exclusive loyalty program that offers rewards, discounts, and exclusive perks based on tenure and usage.
  - Goal: Increase customer lifetime value and create a sense of loyalty.
- **Strategy 4: Bandwidth Guarantee and Rebate:**
  - Action: Offer a bandwidth guarantee with a rebate if service performance consistently falls below the guaranteed levels.
  - Goal: Provide assurance of service quality and demonstrate commitment to customer satisfaction.
- **Strategy 5: Proactive Contract Negotiation & Bundle Optimization:**
  - Action: Reach out to high-value customers before their month-to-month contract expires to discuss potential contract renewals and bundled service options that better meet their evolving needs. Offer compelling incentives for committing to longer-term contracts.
  - Goal: Convert month-to-month customers into long-term subscribers.

C. Month-to-Month at Risk: (avg. tenure 5.3 months, Electronic Check & Credit Card Payment, mix of DSL and Fiber)

- **Strategy 1: Long-Term Contract Conversion Campaign:**
    - Action: Offer significant discounts, price freezes, or bundled service upgrades to incentivize customers on month-to-month plans to switch to longer-term contracts (e.g., 12-month, 24-month).
    - Goal: Reduce churn by locking customers into longer-term commitments.
- **Strategy 2: Automatic Renewal Incentives:**
    - Action: Provide a small, recurring discount or bonus data to customers who enroll in automatic contract renewals.
    - Goal: Encourage customers to stay on current plans without active intervention.
- **Strategy 3: Limited-Time Offers for Longer-Term Subscribers:**
    - Action: Advertise exclusive deals and promotions only available to customers on 12-month or 24-month contracts.
    - Goal: Make the benefits of longer-term contracts more appealing.
- **Strategy 4: Tiered Support Levels Based on Contract Length:**
    - Action: Provide faster and more personalized customer support to customers on longer-term contracts.
    - Goal: Reward customer loyalty with better service.
- **Strategy 5: Early Renewal Benefits:**
    - Action: Offer incentives for customers to renew their contract before it expires, such as additional data or a free month of service.
    - Goal: Reduce uncertainty and encourage early commitment.

D. Fiber Optic at Risk: (avg. tenure 6 months, Credit Card Payment, Month-to-Month)

- **Strategy 1: Root Cause Analysis of Performance Issues:**
    - Action: Conduct a thorough investigation to identify common performance issues experienced by Fiber Optic customers (e.g., congestion, equipment limitations, software problems).
    - Goal: Address underlying technical issues that may be causing dissatisfaction.
- **Strategy 2: Network Optimization and Upgrade Program:**
    - Action: Invest in network upgrades and optimization measures to improve Fiber Optic service reliability, speed consistency, and overall performance.
    - Goal: Deliver the performance that customers expect from a premium Fiber Optic service.
- **Strategy 3: Advanced Troubleshooting Tools and Resources:**
    - Action: Provide Fiber Optic customers with access to advanced troubleshooting tools, self-help resources, and expert technical support to quickly resolve technical issues.
    - Goal: Empower customers to resolve common problems on their own and improve the overall support experience.
- **Strategy 4: Bandwidth Monitoring and Usage Alerts:**
    - Action: Offer bandwidth monitoring tools and usage alerts to help customers understand their data consumption and optimize their service usage.
    - Goal: Prevent unexpected data overages and improve customer satisfaction with service performance.
- **Strategy 5: Personalized Fiber Optic Service Optimization Consultation:**
    - Action: Offer personalized consultations with Fiber Optic service experts to help customers optimize their service configuration, equipment settings, and network setup for maximum performance.
    - Goal: Provide tailored recommendations and support to ensure customers are getting the most out of their Fiber Optic service.

3. Metrics to Track Effectiveness:

To assess the effectiveness of these churn reduction strategies, the following metrics should be monitored:

- **Churn Rate by Segment:** Track the churn rate for each of the identified high-risk segments over time to measure the impact of targeted interventions.
- **Customer Satisfaction (CSAT) Scores:** Measure customer satisfaction scores within each segment through surveys, feedback forms, and online reviews to gauge the effectiveness of customer support and service improvements.
- **Net Promoter Score (NPS):** Track the NPS for each segment to assess customer loyalty and advocacy.
- **Conversion Rates:** Monitor the conversion rates for initiatives such as payment method upgrades and contract renewals to measure the success of these programs.
- **Customer Lifetime Value (CLTV):** Calculate the CLTV for customers in each segment to assess the long-term financial impact of churn reduction efforts.
- **Support Ticket Volume and Resolution Time:** Monitor the volume of support tickets and the

average resolution time for each segment to assess the effectiveness of customer support improvements.

- **Uptake Rate of Proactive Offers:** Track the percentage of customers who accept proactive offers (e.g., discounts, bonus data) to measure the relevance and effectiveness of these incentives.

By implementing these strategies and tracking these metrics, the telecom provider can significantly reduce churn and improve customer loyalty. Remember to continuously monitor and adjust these strategies based on performance and evolving customer needs.

5. FUTURE ENHANCEMENT

While the current model demonstrates high accuracy and scalability in predicting customer churn, several opportunities exist for future enhancements. Firstly, the integration of **real-time data pipelines** using tools like Apache Kafka or Spark Streaming could enable near-instant churn prediction and proactive customer engagement. Secondly, expanding the feature set to include **customer sentiment analysis from call center transcripts or social media interactions** can provide deeper behavioral insights. Additionally, experimenting with advanced algorithms such as **XGBoost, LightGBM, or deep learning models** (e.g., LSTM networks for sequential behavior) may further improve predictive performance, especially in capturing temporal patterns. Incorporating **explainable AI (XAI)** techniques like SHAP values could enhance the interpretability of the model, allowing business teams to better understand individual churn decisions.

Finally, deploying the model as a **scalable API or microservice** within a customer relationship management (CRM) system can facilitate seamless integration with business workflows and drive data-informed retention strategies.

6. CONCLUSION

Finally, this work has shown that machine learning algorithms such as Random Forest, K-Nearest Neighbours (KNN), and Logistic Regression are successful in identifying transactions that involve banking fraud. We have created strong models that can reliably identify transactions as either fraudulent or legitimate by tackling the class imbalance issue with meticulous algorithm selection,

feature enhancement strategies, and thorough assessment criteria. The results emphasise how crucial it is to use a

variety of methods for detecting fraud because every algorithm has unique advantages.

7. REFERENCES

[1] Idris, A., Rizwan, M., & Khan, A. (2012). Churn prediction in telecom using Random Forest and PCA. International Journal of Computer Applications, 27(11), 26-31.

[2] Ahmed, A., & Maheswari, D. (2019). Customer Churn Prediction in Telecom Industry using Logistic Regression and Decision Tree. Procedia Computer Science, 165, 719–725.

[3] Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2012). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications, 38(3), 2354-2364.

[4] Lariviere, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications, 29(2), 472-484.

[5] Singh, D., & Reddy, C. K. (2021). A survey on platforms for big data analytics. Journal of Big Data, 8(1), 1-54.

[6] Shinde, S., & Kulkarni, U. (2020). Machine learning and big data for churn prediction: A comparative review. Journal of Engineering Science and Technology Review, 13(4), 15–23.

[7] Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

[8] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211-229.

[9] Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., Hawalah, A. Y., & Hussain, A. (2019). Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing, 237, 242-254.

[10] Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadi, A., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. Applied Soft Computing, 24, 994-1012.

[11] Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1), 1414-1425.

[12] Ahmed, A., & Maheswari, U. S. (2019). Customer churn prediction in telecom using machine learning in big data platform. International Journal of Engineering and Advanced Technology, 8(6), 1413-1416.

[13] Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1-9.

[14] Sharma, A., & Panigrahi, P. K. (2013). A neural network based approach for predicting customer churn in cellular network services. International Journal of Computer Applications, 27(11), 26-31.

[15] Lariviere, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications, 29(2), 472-484.