iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

TELECOM CUSTOMER CHURN PREDICTION USING MACHINE LEARNING

Karri LakshmiPrasad (MCA Student),
Dr. A. Mary Sowjanya (Associate Professor),
Varsha Kumari (Research Scholar)
Department of Information Technology and Computer Applications
Andhra University College of Engineering, Visakhapatnam, AP.
Corresponding Author: Karri LakshmiPrasad,
Email: prasadkarri0407@gmail.com

ABSTRACT
For telecommunication companies, losing customers can be particularly important since it can impact revenue and future customer relationships. Determining customers who are at risk of leaving early can help companies prevent their customers from leaving them. This project is a machine learning system for predicting customer churn amongst telecom service providers based on the demographic, account, and service related information. It is based on the Telco Customer Churn dataset provided by IBM which has 33 attributes and 7,043 customers records. Unnecessary fields, like customer identifiers and geographic data, were stripped of from the data during preprocessing, and variables that might lead to data leakage (like churn score, churn value, and churn reason) were also removed. All the other data was cleaned with the process of missing value handling and one-hot coding of categorical variables. Numerical features were standardized using Standard Scaler. Logistic Regression was chosen as the main classification model and trained with 80:20 train-test split. The model was tested against 1,409 test customer records and it performed with an accuracy of 80.3%, precision of 67.9%, and a recall of 57.8% on the churn class. The model also achieved an ROC-AUC of 0.854, which indicates that the model has a good capacity to distinguish between customers who are likely to churn and those who are likely to stay. Furthermore, 5-fold stratified cross validation was used to assess the performance of the model, which yielded a mean accuracy of 81.09% and a mean ROC-AUC of 0.858, supporting the performance of the model. The trained model, scaler, and feature information were saved using Joblib to be used practically in the Streamlit web application. The application enables the user to input the customer information, and instantly gets a probability of churn along with a risk indicator. Various functional, interface, edge case, backend and security scenarios were tested. Churn probabilities of 0.58 and 0.01 were returned for example customer profiles representing higher and lower risk customers respectively. The project provides a valuable demonstration of the integration of machine learning, data preprocessing techniques, and an interactive interface to create a practical decision support system for telecom customer retention. The Customer Churn Prediction app is a machine learning application that predicts the churn for customers in the telecom industry. Customer Churn Prediction app is a machine learning application for predicting customer churn in telecom sector.

Keywords:
Customer Churn Prediction, Machine Learning, Logistic Regression, Feature Engineering, Data Preprocessing, Streamlit, Telecom Analytics, Classification, Decision Support System

INTRODUCTION
The telecom industry is facing one of its biggest challenges in customer churn, in which customers can stop services if they are not satisfied with the quality of services, pricing, contract terms, and/or other alternatives available. With the rising competition among telecommunication service providers, customer retention is equally vital as acquiring new customers.. Losing customers can directly affect the revenue and profitability of an organization, while identifying customers who are likely to leave can help service providers take suitable retention measures at an early stage. Therefore, customer churn prediction has become an important application of data analytics and machine learning in the telecommunications domain [1]. The data-driven customer churn prediction framework can derive customer data from the demographic, service usage, billing and contract data to detect patterns correlated with a customer churn. Common attributes found in telecommunication datasets include gender, partner/family member and dependent data, tenure, monthly charges, total charges, internet service, online security, technical support, contract type, paperless billing and payment method. Machine learning algorithms can be used to analyze these characteristics to predict if a customer will churn or not. Machine learning methods have

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

been previously demonstrated to be effective in supporting churn analysis by discovering high-risk customers and understanding factors behind customers' behavior [2]. Factors that impact customer churn may differ, and impact might differ by individual customer characteristics and service conditions. Those who are not customers for long may be more likely to cancel service as they may not have a strong attachment to the service provider. Likewise, the perception of added costs, monthly contracts, absence of extras, and specific payment options could all be linked to greater churn risk. However, customers who have longer service contracts and longer service tenure may be more loyal. Once these patterns are understood, an organization can target its retention initiatives at customers that are more likely to leave [3]. Historical customer data can be used to build predictive models using machine learning, which is an efficient way. The proposed system involves pre-processing customer data, which includes numerical and categorical attributes, transforming categorical data into numerical format, and feature scaling. The logistic regression classification model is then used to estimate the probability of churning. The trained model is embedded within Streamlit based web interface, where the user can input customer details and get a churn prediction results along with its probability. A probability threshold of 0.35 is used to classify customers as likely to churn or not likely to churn. The suggested Telecom Customer Churn Prediction system consequently integrates information preprocessing, machine learning, and a user-friendly web interface to give a usable way of determining potential customer churn. The system can be helpful for organizations to identify high-risk customers and make informed customer retention decisions. The proposed approach is a direct prediction to the customer information, which shows how machine learning can be applied to the telecom industry

OBJECTIVES

Ahmad, Jafar and Aljoumaa (2019) proposed a machine-learning based approach for customer churn prediction in the telecom industry. The study adopted a big data set from SyriaTel and performed feature engineering and feature selection to boost the predictive performance. The researchers also added SNAT features to capture customers relationship and behavior pattern. The proposed approach was assessed based on the main performance measure called Area Under the Curve. The results indicated that the addition of social network features increased the AUC from 84% to 93.3%, proving the value of customer and network-related information in churn prediction in the context of the social network [4]. Saheed and Hambali (2021) put forward a machine learning algorithm-based customer churn prediction method in the telecom industry that utilizes feature selection methods. Four algorithms namely Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest (RF) and Naive Bayes (NB) were evaluated. To select important features prior to model training, the Information Gain and Ranker methods were used. When feature selection was introduced, the accuracy was reported as 95.02% while without feature selection the accuracy was 92.92%. The study also employed 10-fold cross-validation, and demonstrated that feature selection can boost the accuracy of the machine learning models used for customer churn prediction [5].Shrestha and Shakya (2022) proposed a Customer churn prediction model for telecommunication industry in Nepal using XGBoost. They were studying a telecom dataset of their own that was native to their company and included 52,332 customer records, some churned, some not. The researchers concentrated on the unbalancedness of the telecom dataset and assessed the performance of XGBoost model in churn classification. The model performed well with an accuracy of 97% and F1 of 88% on the native dataset. A public dataset was also used, with the model having an accuracy of 96.25% and an F1 score of 86.34% [6]. Similarly, Sikri et al. (2024) explored the problem of predicting customer churn in the telecom sector using various machine learning algorithms. Logistic Regression, Naive Bayes, K-Nearest Neighbour, Decision Tree, Perceptron, Multi-Layer Perceptron, Gradient Boosting and XGBoost were tested and evaluated. To solve class imbalance problem the researchers have applied a technique known as data balancing based on the ratio technique and compared with the existing over sampling and under sampling. They found that ensemble models outperformed some of the individual models in the telecom sector for churn prediction, specifically Gradient Boosting and XGBoost models [7]. Likewise, the importance of preprocessing and suitable classification methods for customer retention has been highlighted in recent works. Previous studies have shown that customer demographic information, service use, contract information, payment information, and billing details can be useful in determining the likelihood of churning customers. Although advanced ensemble models often provide high predictive performance, Logistic Regression remains a useful classification technique because of its simplicity, computational efficiency, and interpretability. Based on these observations, the present project develops a telecom customer churn prediction system using Logistic Regression and provides a Streamlit-based interface for obtaining churn probability from customer information [8].

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

Data Collection

The study uses the IBM Telco Customer Churn dataset containing 7,043 customer records and 33 original attributes. The records include demographic, account, service, contract, payment, and billing information. The target variable is Churn Label, which indicates whether the customer left the service during the observed period. In the project dataset, 5,174 customers are labelled as non-churners and 1,869 as churners, giving a churn proportion of approximately 26.54%.

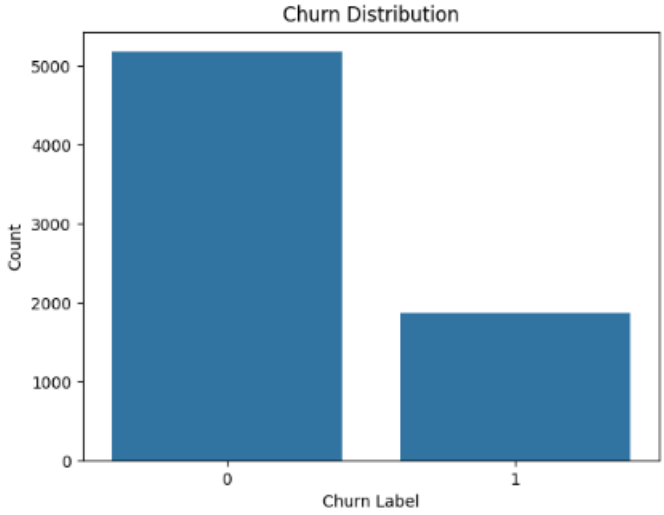

Figure 1: Churn Distribution in the Dataset

Data Preprocessing

Several attributes that are not appropriate as predictive inputs were removed before model development. These included customer identifiers, geographic fields, reporting fields, churn reason, churn score, churn value, and CLTV. Total Charges was converted into numeric form and missing numeric values were handled using the median. The target Churn Label was encoded as 1 for Yes and 0 for No.

Categorical variables were converted into numerical representations using one-hot encoding. After preprocessing, 46 predictor features were obtained. StandardScaler was then applied to standardize the feature matrix. The resulting data was used for the 80:20 train-test evaluation and for stratified five-fold cross-validation.

Machine Learning Model Development

Logistic Regression was selected as the main classification model because the target variable represents a binary churn outcome and the model provides a direct probability estimate. The implementation used the Scikit-learn LogisticRegression classifier with the liblinear solver and a maximum of 200 iterations. A fixed random state of 42 was used to make the experimental split reproducible.

Parameter	value
Classification Algorithm	Logistic Regression
Solver	liblinear
Maximum Iterations	200
Training / Testing Split	80% / 20%
Random State	42
Input Features	46
Cross Validation	5-fold Stratified

Model Validation

The test set contained 1,409 records. Accuracy, precision, recall, F1-score and ROC-AUC were used to evaluate the model. In addition, stratified five-fold cross-validation was performed to examine whether the observed performance was consistent across different data partitions. The cross-validation results reported in the project were a mean accuracy of 81.09% and a mean ROC-AUC of 0.858.

Model and Artifact Saving

After training, the model was saved using Joblib together with the StandardScaler and the final feature-column configuration. These artifacts allow the Streamlit application to reuse the trained model and apply the same feature representation during prediction without retraining the classifier.

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

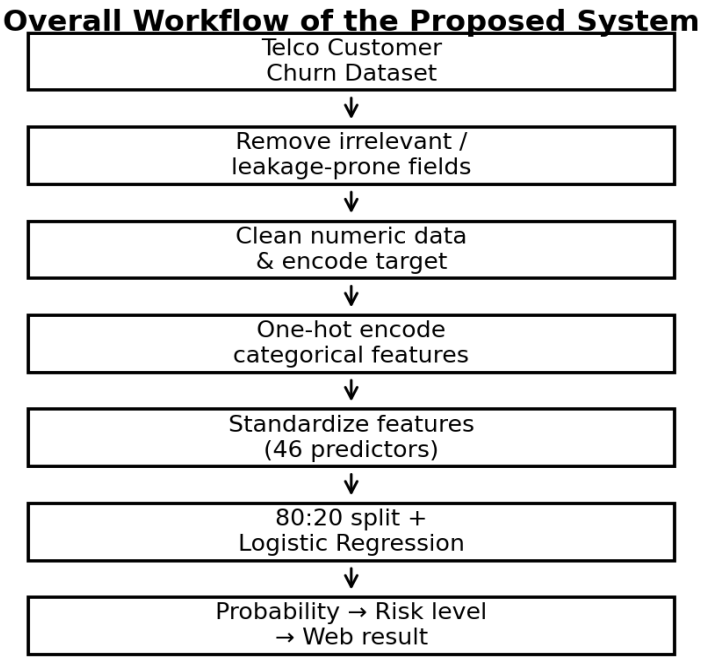

Figure 2: Overall Workflow of the Proposed Telecom Customer Churn Prediction System

WEB-BASED TELECOM CUSTOMER CHURN PREDICTION SYSTEM

A Streamlit-based web application was developed to provide direct access to the trained churn prediction model. The application contains three main areas: Single Prediction, Batch Prediction (CSV), and About the Model. The interface loads the saved model artifacts and performs the prediction using the same feature structure used during model development.

Single Customer Prediction

The Single Prediction module allows a user to enter customer demographic, account, service, contract, and payment information. The entered values are converted into the encoded feature representation expected by the trained model. The application then applies the saved scaler and obtains the churn probability using predict_proba. The result is displayed together with a risk indicator and a short recommendation.

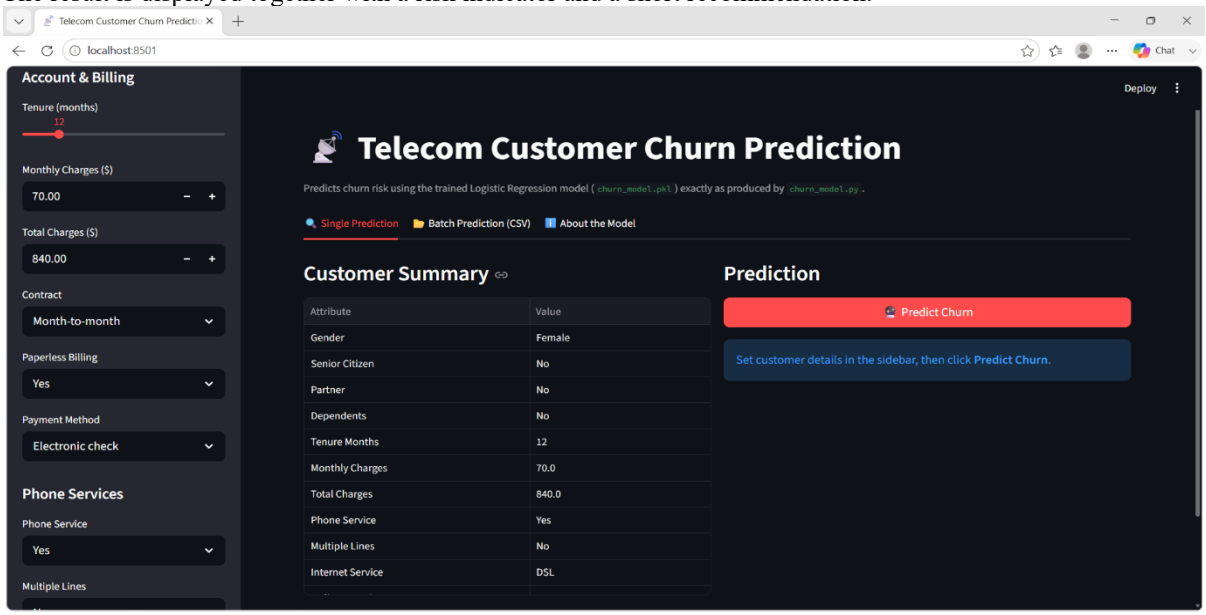

Figure 3: Single Customer Prediction Interface

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

Batch Customer Prediction

The Batch Prediction module accepts a CSV file containing the customer attributes required by the model. The uploaded records are transformed using the same feature encoding and column-alignment procedure as the single prediction module. The application returns the churn probability and risk level for each record and provides the resulting data as a downloadable CSV file.

Churn Risk Classification

The application converts the predicted probability into an operational risk indicator. A probability below 0.35 is treated as Low Risk, a probability from 0.35 to below 0.60 is treated as Medium Risk, and a probability of 0.60 or higher is treated as High Risk. The 0.35 threshold is also used as the main binary decision boundary when a customer is considered likely to churn or not likely to churn.

Churn Probability	Risk Level	Suggested Interpretation
<0.35	Low Risk	Customer is less likely to churn
0.35 – < 0.60	Medium Risk	Customer may require monitoring
≥ 0.60	High Risk	Consider proactive retention outreach

SYSTEM DESIGN AND UML REPRESENTATION

The system design describes how the user interacts with the prediction application and how the customer information moves through the preprocessing and machine learning components. The diagrams are based on the implemented Streamlit application and the saved model ar**tifacts.**

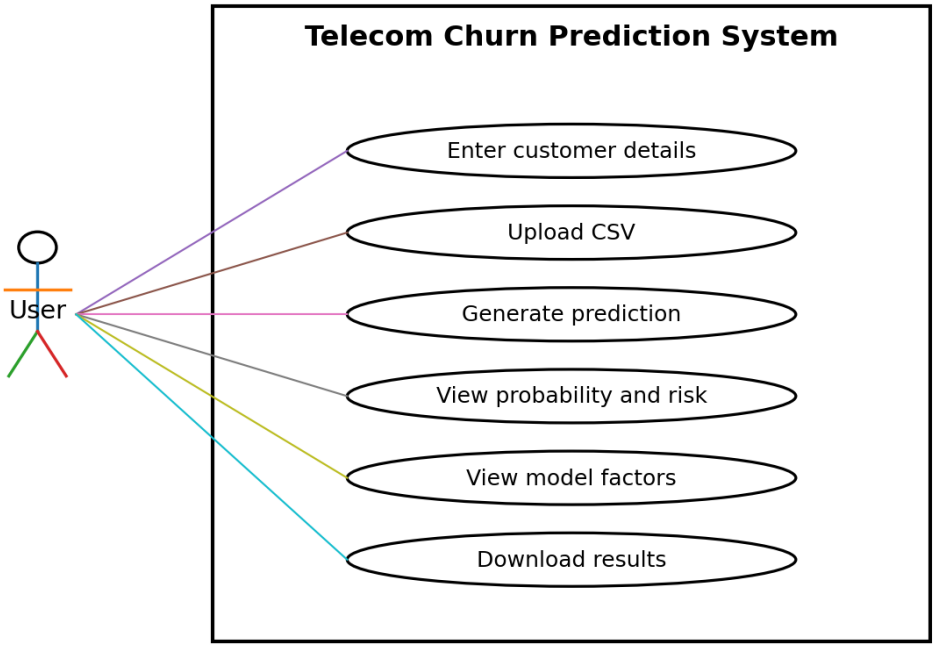

Figure 4: Use Case Diagram of the Telecom Customer Churn Prediction System

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

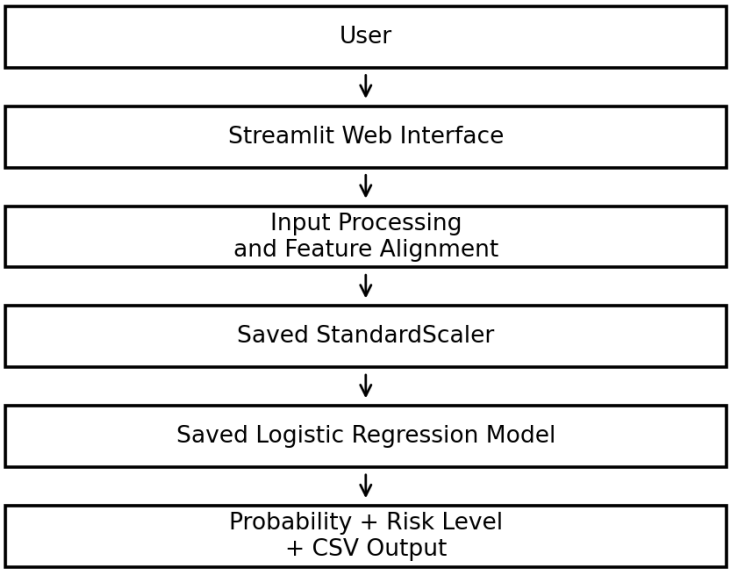

Figure 5: System Architecture of the Proposed Telecom Customer Churn Prediction System

The use case diagram represents the main functions available to the user, including entering customer information, uploading a CSV file, generating a prediction, viewing the probability and risk level, reviewing the model factors, and downloading results. The activity diagram shows the operational sequence from input collection to prediction. The architecture diagram separates the user interface, feature-processing stage, saved preprocessing artifact, trained classifier, and final output.

RESULTS AND DISCUSSION

The developed Logistic Regression model was evaluated using the held-out test data and stratified five-fold cross-validation. The results show that the model provides a useful baseline for separating churn and non-churn customers while maintaining a simple and interpretable classification approach

Confusion Matrix of the Model

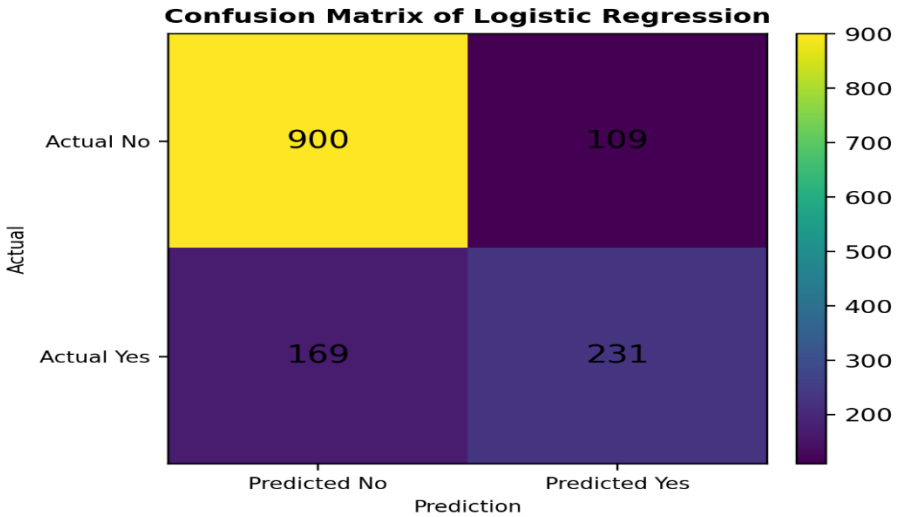

Figure 6: Confusion Matrix of the Logistic Regression Model

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

The confusion matrix contains 900 correctly classified non-churning customers and 231 correctly classified churners. There were 109 non-churning customers classified as churners and 169 churners classified as non-churners. The false-negative cases are particularly important in a retention setting because they represent customers who could potentially leave without being identified by the prediction system.

Classification Performance

Metric	Test Set	5-Fold Stratified CV	
Accuracy	80.27%	81.09% mean	
Precision	67.94%	—	
Recall	57.75%	—	
F1-Score	62.43%	—	
ROC-AUC	0.854	0.858 mean	

The model achieved an accuracy of 80.27% on the test set, with precision of 67.94%, recall of 57.75%, and an F1-score of 62.43%. The ROC-AUC of 0.854 indicates useful discrimination between churn and non-churn customers. The mean five-fold cross-validation accuracy of 81.09% and mean ROC-AUC of 0.858 are close to the held-out test results, providing additional support for the consistency of the model's performance.

At the same time, the recall of 57.75% shows that a considerable number of actual churners were not detected. Therefore, the system is more appropriately viewed as a decision-support and customer-prioritization tool rather than as an autonomous system that should make retention decisions without additional business information.

PREDICTION RESULTS THROUGH THE WEB APPLICATION

The web application was tested with different customer profiles to verify the end-to-end prediction process. The interface accepts customer information, applies the saved preprocessing components, and displays the resulting probability and risk category.

High-Risk Prediction Example

One demonstrated customer profile produced a churn probability of 79.7% and was classified as High Risk. The interface recommends proactive retention outreach for this category. The result represents the model's estimated probability for the entered profile and should not be interpreted as a guarantee that the customer will churn.

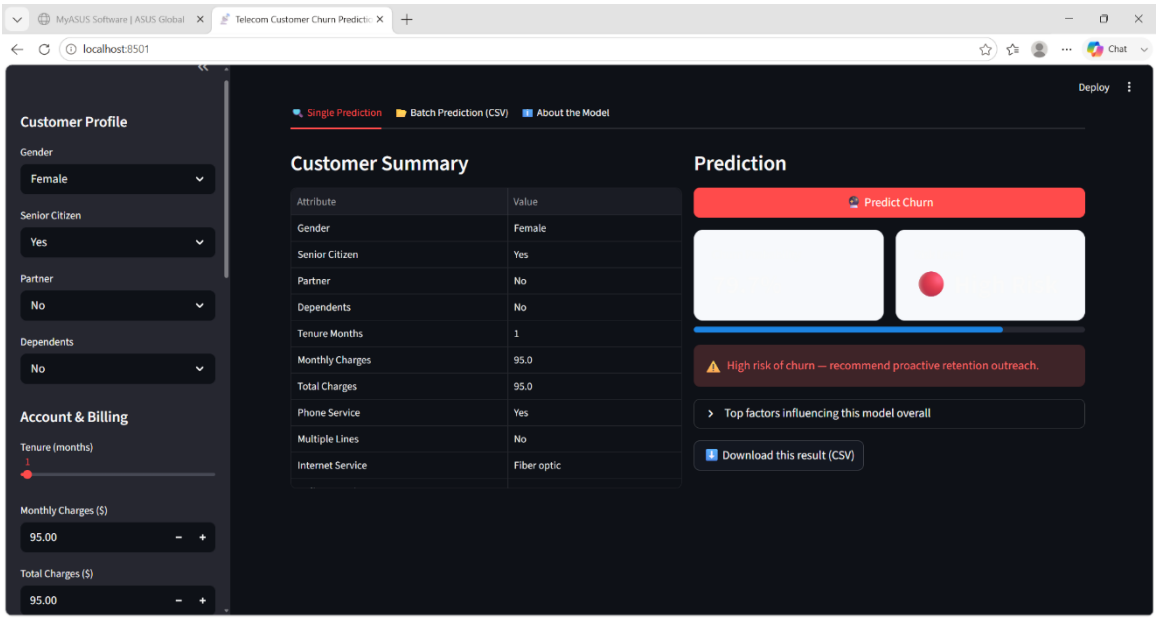

Figure 7: High-Risk Customer Prediction

Low-Risk Prediction Example

Another demonstrated customer profile produced a churn probability of 27.2% and was classified as Low Risk. The application displayed a message indicating that the customer was unlikely to churn. These examples illustrate how the probability output is translated into a simple risk category for practical use.

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

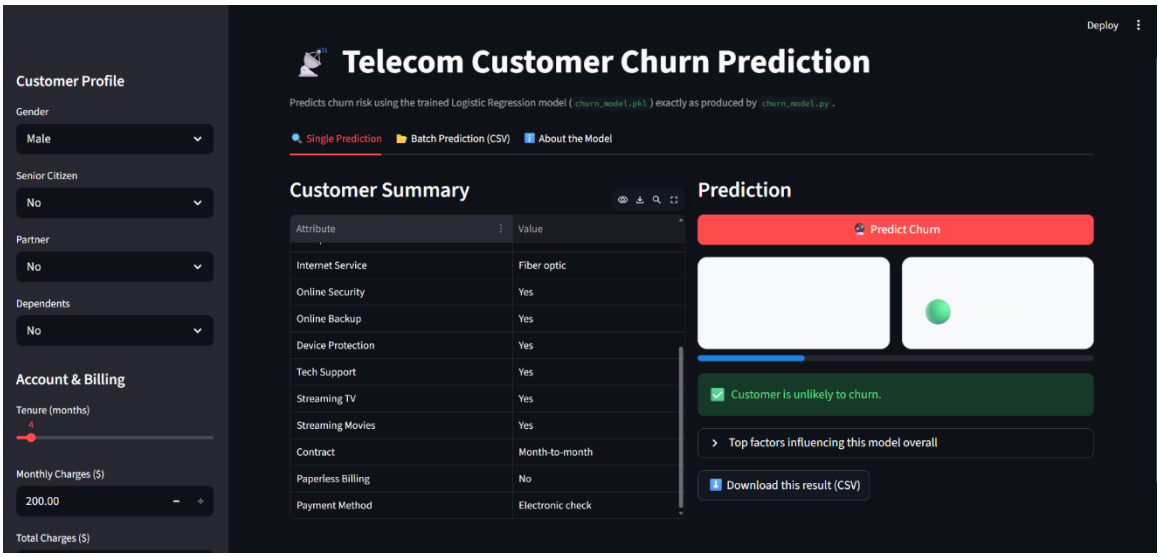

Figure 8: Low-Risk Customer Prediction

Model Factors

The application also provides global Logistic Regression coefficients to give users a simple view of the features that have stronger influence on the trained model. These coefficients describe the model globally and are not individualized explanations for a particular customer.

Feature	Coefficient	Direction	
Tenure Months	-1.4083	Negative	
Total Charges	+0.6461	Positive	
Monthly Charges	-0.3360	Negative	
Dependents	±0.3199	Encoded pair	
Contract – Two year	-0.3155	Negative	
Contract – Month-to-month	+0.3014	Positive	
Internet Service – Fiber optic	+0.2585	Positive	
Internet Service – DSL	-0.2142	Negative	

SYSTEM TESTING AND DISCUSSION

The implemented application was checked through functional, interface, edge-case, backend and security-oriented scenarios. The functional checks focused on prediction generation and CSV processing, while interface checks examined the display of probability and risk indicators. Edge cases included customer inputs at different tenure and charge levels and the use of categorical combinations supported by the application. Backend checks verified that the saved model, scaler and feature-column configuration were loaded correctly during prediction.

The overall results indicate that the project successfully connects a conventional machine learning classifier with an interactive application. The main practical benefit is that the user does not need to run the training code directly to obtain a prediction. The batch mode also makes the system more suitable for screening a group of customers, while the risk categories provide a simple way to prioritize further attention.

There are limitations that should be considered when interpreting the reported results. The project evaluates Logistic Regression as the implemented predictive model and does not provide a controlled comparison with the imported Decision Tree or Random Forest classes, which were not trained or evaluated in the project code. Hyperparameter optimization and external validation were also not performed. In addition, the implementation standardizes the full feature matrix before the train-test split; therefore, the reported test and cross-validation results should be interpreted cautiously because preprocessing leakage may introduce a small optimistic bias. A future implementation should place preprocessing inside a cross-validation pipeline and use a leakage-safe stratified evaluation procedure.

SYSTEM TESTING AND DISCUSSION

The implemented application was checked through functional, interface, edge-case, backend and security-oriented scenarios. The functional checks focused on prediction generation and CSV processing, while interface checks

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

examined the display of probability and risk indicators. Edge cases included customer inputs at different tenure and charge levels and the use of categorical combinations supported by the application. Backend checks verified that the saved model, scaler and feature-column configuration were loaded correctly during prediction.

The overall results indicate that the project successfully connects a conventional machine learning classifier with an interactive application. The main practical benefit is that the user does not need to run the training code directly to obtain a prediction. The batch mode also makes the system more suitable for screening a group of customers, while the risk categories provide a simple way to prioritize further attention.

There are limitations that should be considered when interpreting the reported results. The project evaluates Logistic Regression as the implemented predictive model and does not provide a controlled comparison with the imported Decision Tree or Random Forest classes, which were not trained or evaluated in the project code. Hyperparameter optimization and external validation were also not performed. In addition, the implementation standardizes the full feature matrix before the train-test split; therefore, the reported test and cross-validation results should be interpreted cautiously because preprocessing leakage may introduce a small optimistic bias. A future implementation should place preprocessing inside a cross-validation pipeline and use a leakage-safe stratified evaluation procedure.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to the Department of Information Technology and Computer Applications, Andhra University College of Engineering, Visakhapatnam, for providing the academic environment and support required to carry out this project. The authors also thank their faculty members and colleagues for their guidance and valuable suggestions during the development of the system.

REFERENCES

[1] S. A. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. H. Mason, “Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models,” Journal of Marketing Research, vol. 43, no. 2, pp. 204–211, 2006, doi: 10.1509/jmkr.43.2.204.

[2] W. Verbeke, D. Martens, C. Mues, and B. Baesens, “Building comprehensible customer churn prediction models with advanced rule induction techniques,” Expert Systems with Applications, vol. 38, no. 3, pp. 2354–2364, 2011, doi: 10.1016/j.eswa.2010.08.023.

[3] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, “New insights into churn prediction in the telecommunication sector: A profit driven data mining approach,” European Journal of Operational Research, vol. 218, no. 1, pp. 211–229, 2012, doi: 10.1016/j.ejor.2011.09.031.

[4] A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform,” Journal of Big Data, vol. 6, Art. no. 28, 2019, doi: 10.1186/s40537-019-0191-6.

[5] Y. K. Saheed and M. A. Hambali, “Customer churn prediction in telecom sector with machine learning and information gain filter feature selection algorithms,” in Proc. 2021 International Conference on Data Analytics for Business and Industry (ICDABI), 2021, pp. 208–213, doi: 10.1109/ICDABI53623.2021.9655792.

[6] S. M. Shrestha and A. Shakya, “A Customer Churn Prediction Model using XGBoost for the Telecommunication Industry in Nepal,” Procedia Computer Science, vol. 215, pp. 652–661, 2022, doi: 10.1016/j.procs.2022.12.067.

[7] A. Sikri, R. Jameel, S. M. Idrees, and H. Kaur, “Enhancing customer retention in telecom industry with machine learning driven churn prediction,” Scientific Reports, vol. 14, Art. no. 13097, 2024, doi: 10.1038/s41598-024-63750-0.

[8] “Customer churn prediction in telecom sector using machine learning techniques,” Reliability Engineering & System Safety, vol. 14, Art. no. 100342, 2024, doi: 10.1016/j.rico.2023.100342.

[9] IBM Cognos Analytics, “Telco Customer Churn” sample dataset, IBM base samples, 2019.

 © Author(s) | CC BY 4.0 International License.