

Churn Guard Pro: A Predictive Modeling of Credit Card Customer Churn Using Machine Learning

Hettiarachchi H.A.S.H

Department of Computer Engineering

Faculty of Engineering

University of Sri Jayewardenepura

Ratmalana, Sri Lanka

en93839@sjp.ac.lk

Abstract—For credit card firms, client churn—the phenomenon of customers stopping their card services—has grown to be a serious problem. It impedes long-term profitability and creates uncertainty about future revenues. This study suggests a data-driven method for machine learning-based customer attrition prediction in order to address this problem. Through the use of both exploratory and descriptive data analysis techniques, a dependable machine learning model is created that can precisely identify clients who are at risk by identifying patterns and relationships within customer data. Using F1 scores, the performance of several machine learning algorithms—Random Forest, Ada-Boost, SVM, and XGBoost is assessed; XGBoost performs better than the others. The final model is deployed in a user-friendly Streamlit-based web application, enabling credit card companies to proactively identify and retain high-value customers. This study advances the tactics used by credit card companies to retain customers by providing a workable way to reduce customer attrition and improve credit stability.

Index Terms—customer churn, machine learning, XGBoost

I. INTRODUCTION

There is fierce competition among credit card firms as they compete to draw in and keep clients. Customer retention is critical for long-term growth and profitability in this dynamic economy. However, because so many people move providers or stop using credit cards completely, credit card firms have a difficult time keeping their clientele. Customer churn is a phenomenon that seriously jeopardizes the financial viability and long-term prospects of credit card firms.

A 2022 study by Javelin Strategy and Research states that losing a credit card account typically costs 236 USD. This cost is a result of three factors: revenue loss, customer acquisition, and churn-related administrative costs. The fact that it is sometimes more expensive to acquire new customers than to keep existing ones exacerbates the effects of customer attrition. The fact that it is sometimes more expensive to acquire new customers than to keep existing ones exacerbates the effects of customer attrition.[1]

The Extreme Gradient Boosting (XGBoost) algorithm is utilized in the machine learning model that was created in this study. A well-liked machine learning algorithm called XGBoost has proven to be very successful in a variety of predictive modeling applications. The XGBoost model was chosen because it performed better than Random Forest, Ad-

aBoost, and Support Vector Machine (SVM) models, among other machine learning models.

The web application created in this study makes use of the open-source Python web application library Streamlit. Streamlit makes it possible to create intuitive web apps with little to no coding knowledge. Credit card companies can input customer data into the web application to receive real-time predictions of customer churn risk through an easy-to-use interface.

Two things this research has contributed are:

A. *The creation of a very precise machine learning model to forecast credit card customer attrition*

B. *The creation of an intuitive online application that makes use of the machine learning model to identify clients who are at risk*

This research has the potential to significantly impact the credit card business by giving credit card companies an effective tool to improve customer retention and lower customer churn.

The suggested strategy has the potential to increase customer satisfaction and loyalty, in addition to its financial advantages, by giving credit card companies the ability to proactively handle customer complaints and improve their overall customer experience.

II. RELATED WORK

The practice of determining which customers are most likely to end their commercial relationship with a company—known as customer churn prediction—has drawn a lot of interest recently since it has a big impact on both customer satisfaction and corporate profitability. Algorithms for machine learning (ML) have become extremely useful and accurate methods for forecasting client attrition.

In order to forecast credit card user churn, Vafeiadis et al. [3] examined the effectiveness of four machine learning (ML) algorithms: logistic regression, decision trees, random forests, and support vector machines (SVMs). SVMs proved to be a suitable choice for churn prediction jobs as they performed better than the other algorithms in terms of accuracy and F-measure.

Building on this work, Diamantaras et al. [4] gave an extensive review of machine learning methods for customer churn prediction, exploring the advantages and disadvantages of several algorithms and providing suggestions for choosing the best one for a given set of circumstances. The study reaffirmed the effectiveness of machine learning methods in anticipating client attrition.

Together, these studies demonstrate how well ML approaches work to forecast customer attrition, giving organizations insightful information about the factors that influence attrition and the ability to target interventions to keep consumers.

III. OBJECTIVES

- Conduct descriptive and exploratory data analysis of customer churn data.
- Develop a reliable machine learning model to make predictions of customer churn
- Deployment of the model and creation of the User interface

IV. METHADODOLOGY

This study develops a predictive model for credit card user churn prediction using an extensive data-driven approach. Data collection, preprocessing, exploratory and descriptive data analysis, machine learning model construction, and web application deployment are all included in the framework.

The process of collecting data includes obtaining pertinent client information from a trustworthy source. In this instance, the Kaggle website provided the dataset. Cleaning, dealing with missing values, and formatting data so that machine learning algorithms may use it are all included in data preprocessing.

The goal of descriptive and exploratory data analysis is to find links, patterns, and trends in the data. Understanding consumer behavior and churn patterns is made possible by this study, and it is essential for creating a predictive model that works.

The process of developing a machine learning model includes choosing and developing suitable machine learning algorithms to recognize clients who are at risk. Based on performance measures, several methods were assessed, such as support vector machines, decision trees, random forests, and logistic regression.

Finally, Stream lit was used to create an intuitive web application that deployed the predictive model, making it easy for credit card firms to recognize and keep at-risk consumers. The web application offers a user-friendly interface for entering client information and obtaining up-to-date estimates of the risk of customer attrition.

A. Data Collection

The Kaggle website provided the data used in this study. Information on 10,000 credit card users worldwide is included in the collection. It has 23 features that cover a broad spectrum of transaction patterns, churn status, and customer demographics.

Data columns (total 23 columns):

#	Column	Non-Null Count	Dtype
0	CREDITLIMIT	10127 non-null	int64
1	Attrition_Flag	10127 non-null	object
2	Customer_Age	10127 non-null	int64
3	Gender	10127 non-null	object
4	Dependent_count	10127 non-null	int64
5	Education_Level	10127 non-null	object
6	Marital_Status	10127 non-null	object
7	Income_Category	10127 non-null	object
8	Card_Category	10127 non-null	object
9	Months_on_book	10127 non-null	int64
10	Total_Relationship_Count	10127 non-null	int64
11	Months_Inactive_12_mon	10127 non-null	int64
12	Contacts_Count_12_mon	10127 non-null	int64
13	Credit_Limit	10127 non-null	float64
14	Total_Revolving_Bal	10127 non-null	int64
15	Avg_Open_To_Buy	10127 non-null	float64
16	Total_Amt_Chng_Q4_Q1	10127 non-null	float64
17	Total_Trans_Amt	10127 non-null	int64
18	Total_Trans_Ct	10127 non-null	int64
19	Total_Ct_Chng_Q4_Q1	10127 non-null	float64
20	Avg_Utilization_Ratio	10127 non-null	float64
21	Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1	10127 non-null	int64
22	Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2	10127 non-null	float64

dtypes: float64(7), int64(16), object(6)

Fig. 1. Data set attributes

B. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process that involves investigating and summarizing data to understand its characteristics, patterns, and relationships. It serves as a foundation for further analysis, modeling, and decision-making.

In the EDA for this research it includes techniques, such as:

- Data cleaning: identification and handling of missing values, outliers, and inconsistencies to ensure data integrity.
- Descriptive statistics: calculation of measures of central tendency (mean, median, mode) and dispersion (range, variance, standard deviation) to glean insights into the overall distribution of variables.
- Data visualization: making charts and graphs to show data distributions, identify trends and patterns, and reveal correlations between variables.

1) *Descriptive statistics:* To learn more about the traits of the credit card client base, a preliminary descriptive analysis of the dataset was carried out. For each variable in this analysis, summary statistics comprising the mean, median, standard deviation, minimum, and maximum values were calculated. The results of the descriptive analysis revealed several key

	count	mean	std	min	25%	50%	75%	max
Customer_Age	10127.0	46.325960	8.016614	26.0	41.000	46.000	52.000	73.000
Dependent_count	10127.0	2.346203	1.298908	0.0	1.000	2.000	3.000	5.000
Months_on_book	10127.0	35.928409	7.986416	13.0	31.000	36.000	40.000	56.000
Total_Relationship_Count	10127.0	3.612580	1.554408	1.0	3.000	4.000	5.000	6.000
Months_Inactive_12_mon	10127.0	2.341167	1.010622	0.0	2.000	2.000	3.000	6.000
Contacts_Count_12_mon	10127.0	2.455317	1.106225	0.0	2.000	2.000	3.000	6.000
Credit_Limit	10127.0	8631.953698	9088.776650	1438.3	2555.000	4549.000	11067.500	34516.000
Total_Revolving_Bal	10127.0	1162.814061	814.987335	0.0	359.000	1276.000	1784.000	2517.000
Avg_Open_To_Buy	10127.0	7469.139637	9090.685324	3.0	1324.500	3474.000	9859.000	34516.000
Total_Amt_Chng_Q4_Q1	10127.0	0.759941	0.219207	0.0	0.631	0.736	0.859	3.397
Total_Trans_Amt	10127.0	4404.086304	3397.129254	510.0	2155.500	3899.000	4741.000	18484.000
Total_Trans_Ct	10127.0	64.658695	23.472570	10.0	45.000	67.000	81.000	139.000
Total_Ct_Chng_Q4_Q1	10127.0	0.712222	0.238086	0.0	0.582	0.702	0.818	3.714
Avg_Utilization_Ratio	10127.0	0.274894	0.275691	0.0	0.023	0.176	0.503	0.999

Fig. 2. descriptive analysis of the dataset

characteristics of the credit card customer population:

The average customer age was 46.3 years, with a standard deviation of 8.02 years. The majority of customers had two or three dependents (mean = 2.35, standard deviation = 1.30). Customers had an average of 35.9 months of tenure with the

company (standard deviation = 7.99). They had an average of 3.81 total relationship counts (standard deviation = 1.55) and an average of 2.34 inactive months in the past 12 months (standard deviation = 1.01). Customers had an average of 2.45 contacts with the company in the past 12 months (standard deviation = 1.11). Their average credit limit was \$8,632

2) *Data visualization*: A visual analysis of the data was done in order to find probable trends connected to customer churn and to further investigate the features of the credit card client demographic. As part of this investigation, important variable distributions for both retained and churned consumers were visualized. The distribution of the customer's age, card type, gender, education level, marital status, and credit limit was revealed by the histograms, bar charts, and pie charts that were included in the visualizations.

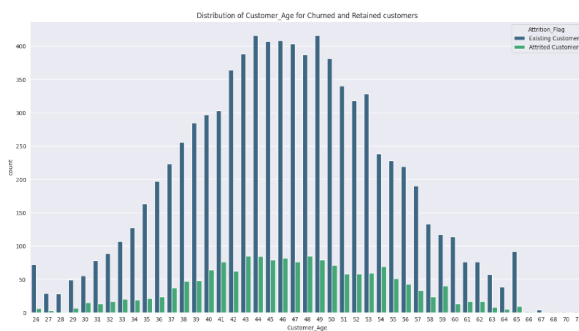


Fig. 3. Distribution of Customer Age for Churned and Retained Customers

The histogram shows that the distribution of customer age is similar for churned and retained customers, with a slight concentration of churned customers in the younger age groups (26-30 years old). However, the overall distribution is fairly even across all age groups, suggesting that customer age is not a major predictor of customer churn.

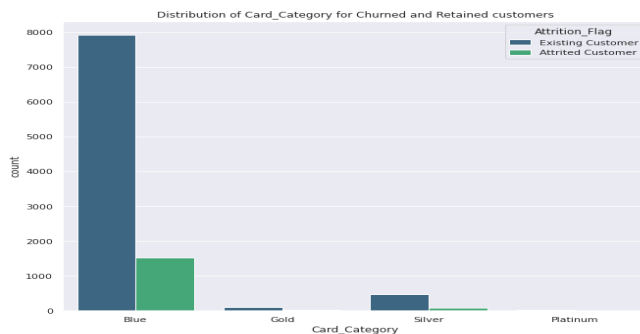


Fig. 4. Distribution of card categories for Churned and Retained customers

C. Data Pre Processing

- Data cleaning: identification and handling of missing values, outliers, and inconsistencies to ensure data integrity.
- Data up-sampling: SMOTE (Synthetic Minority Over-sampling Technique) is a technique used for data up-

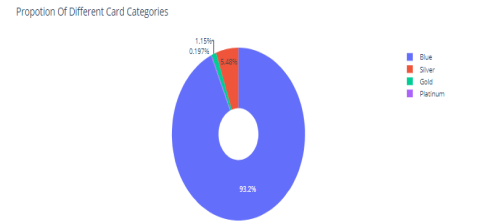


Fig. 5. proportion of card categories

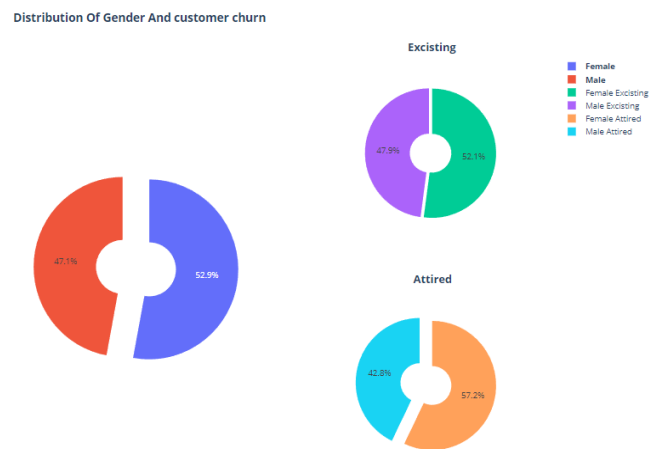


Fig. 6. Distribution of the Genders for Churned and Retained customers

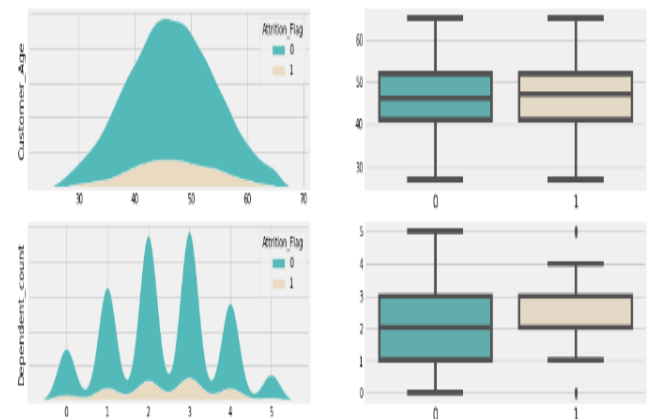


Fig. 7. Exploring Numerical Features

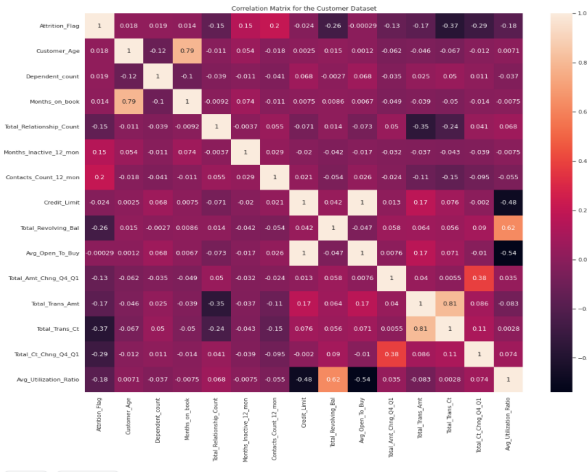


Fig. 8. Correlation Matrix

sampling in the context of imbalanced data sets. It generates synthetic samples for the minority class to balance the class distribution.

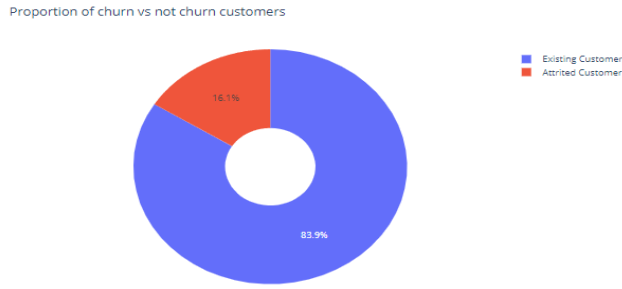


Fig. 9. Data Distribution

D. Development of the Machine Learning models for the data set

Four different machine learning pipelines—Random Forest, XGBoost, Adaboost, and Support Vector Classification (SVC) were built and assessed using ten-fold cross-validation in order to determine which classification technique is best for forecasting customer attrition. A thorough data pre-processing step that included categorical variable encoding, numerical variable normalization, and missing value imputation was applied to each pipeline. Feature selection approaches were then used to determine which features were most relevant for churn prediction, which reduced computational load and improved prediction accuracy.

After a comparative analysis of the four pipelines, XGBoost was shown to be the most reliable classifier, outperforming the others in terms of accuracy, precision, recall, and F1-score. Random Forest followed closely behind, exhibiting an admirable ability for churn prediction. Even though Adaboost and SVC had poorer accuracy, they nevertheless offered insightful information about the dynamics of churn prediction.

TABLE I
COMPARISON OF PERFORMANCE OF MACHINE LEARNING MODELS FOR CUSTOMER CHURN PREDICTION

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.91	0.92	0.9	0.91
XGBoost	0.92	0.92	0.9	0.91
Adaboost	0.87	0.87	0.88	0.87
SVC	0.88	0.89	0.86	0.87

These results highlight how well Random Forest and XGBoost perform in this data set's context when it comes to predicting customer attrition. Their exceptional performance is a result of their capacity to manage intricate variable interactions, efficiently traverse high-dimensional data sets, and reduce over-fitting.

E. Final Model Development

After obtaining the result from cross-validation, the XGB classifier was selected as the final model for the prediction system. XGB classifier was Trained again with these features

```
<class 'pandas.core.frame.DataFrame'>
Index: 8101 entries, 9866 to 7278
Data columns (total 30 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Customer_Age                             8101 non-null   int64
1   Dependent_count                          8101 non-null   int64
2   Months_on_book                           8101 non-null   int64
3   Total_Relationship_Count                 8101 non-null   int64
4   Months_inactive_12_mon                   8101 non-null   int64
5   Contacts_Count_12_mon                    8101 non-null   float64
6   Credit_Limit                             8101 non-null   float64
7   Total_Revolving_Bal                      8101 non-null   int64
8   Total_Trans_Amt                          8101 non-null   int64
9   Total_Trans_Ct                           8101 non-null   int64
10  Card_Category_Blue                       8101 non-null   bool
11  Card_Category_Gold                       8101 non-null   bool
12  Card_Category_Platinum                   8101 non-null   bool
13  Card_Category_Silver                     8101 non-null   bool
14  Income_Category_120K - $50K             8101 non-null   bool
15  Income_Category_50K - $80K              8101 non-null   bool
16  Income_Category_80K - $120K             8101 non-null   bool
17  Income_Category_Less than $50K          8101 non-null   bool
18  Gender_F                                  8101 non-null   bool
19  Gender_M                                  8101 non-null   bool
20  Marital_Status_Divorced                  8101 non-null   bool
21  Marital_Status_Married                   8101 non-null   bool
22  Marital_Status_Single                    8101 non-null   bool
23  Education_Level_College                  8101 non-null   bool
24  Education_Level_Doctorate                8101 non-null   bool
25  Education_Level_Graduate                 8101 non-null   bool
26  Education_Level_High_School              8101 non-null   bool
27  Education_Level_Post-graduate            8101 non-null   bool
28  Education_Level_Uneducated               8101 non-null   bool
29  dtypes: bool(20), float64(1), int64(9)
```

Fig. 10. Training Attributes

	precision	recall	f1-score	support
0	0.97	0.98	0.97	1699
1	0.89	0.84	0.86	327
accuracy			0.96	2026
macro avg	0.93	0.91	0.92	2026
weighted avg	0.96	0.96	0.96	2026

Fig. 11. XGB model results

Training With these input features, the XGB model could achieve accuracy scores as shown in Figure 11.

XGB model architecture can be shown as follows:

In terms of predicting customer attrition, the XGBoost model performed remarkably well, with an overall accuracy of 95.7%, a precision of 96.9%, and a recall of 97.9%. These numbers show that the model has a high degree of accuracy in classifying both retained and churned clients.

The ability of the model to differentiate between retained and churned consumers was further supported by the ROC curve study, which showed an AUC (area under the curve) of

```

XGBClassifier
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric=None, feature_types=None,
              gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=None, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=None, max_leaves=None,
              min_child_weight=None, missing=None, monotone_constraints=None,
              n_estimators=100, n_jobs=-1, num_parallel_tree=None,
              predictor=None, random_state=69, ...)

```

Fig. 12. XGB model

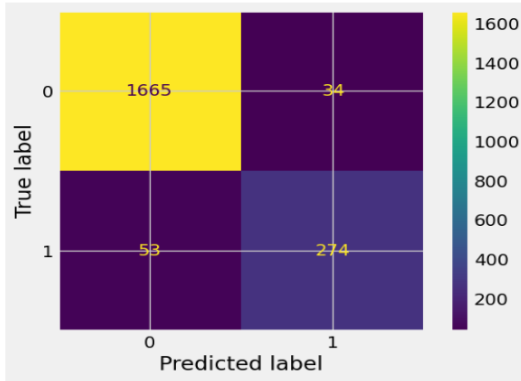


Fig. 13. Confusion Matrix of XGB model

0.99. The model can efficiently distinguish between the two classes even when their distributions overlap, as seen by the high AUC value.

With a recall of 97.9%, the model performed especially well in identifying customers who had churned. This shows that the model can identify the trends that cause customers to leave, which makes proactive intervention options possible.

To sum up, the XGBoost model has proven to be remarkably effective at predicting customer attrition, as demonstrated by its high scores for accuracy, precision, recall, and AUC. Although it has a little bias towards false positives, its efficacy in detecting both retained and churned consumers is consistent with the practical objective of reducing customer attrition. These results imply that the XGBoost model is a useful model to be deployed in a prediction system that aims to predict in order to reduce customer churn in credit card companies

V. WEB APPLICATION

A. Development of Web App

To provide a user-friendly interface for interacting with the trained XGBoost churn prediction model, a web application was developed using the Streamlit framework. A Python package called Streamlit makes it easier to create interactive web apps without needing a lot of experience with front-end development. It smoothly combines server-side (back end) and front-end (UI) features into one Python environment.

Because of the online application's dynamic input handling characteristics, users can supply numerical and categorical inputs that correspond to the attributes that the XGBoost

model uses. Because of the input's versatility, users can provide a wealth of data, which guarantees that the model will have access to all the information it needs to predict churn accurately.

An SQLite database was used to create a basic user authentication system that improved user privacy and security. To guarantee that only authorized users can access the web application and its prediction features, this system enforces secure login and registration procedures.

To preserve user state tracking, the web application additionally makes use of session state management. This functionality guarantees that user interactions and inputs are retained across the application, resulting in a smooth and customized user experience.

In general, the created online application functions as a safe and easy-to-use interface for communicating with the XGBoost churn prediction model. The application's usability and efficacy are improved by its dynamic input handling, user authentication, and session state management features. By incorporating the trained model into the web application, customers can now get real-time churn predictions and take preventative action against possible client attrition.

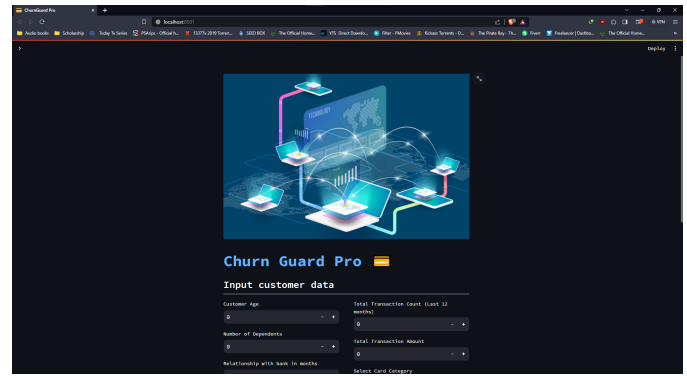


Fig. 14. Web application

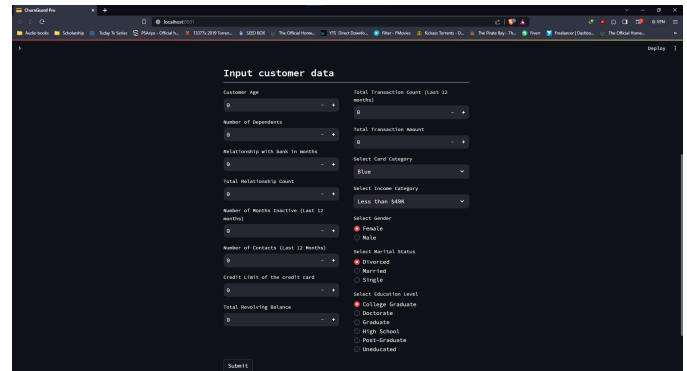


Fig. 15. Web application Contd

B. Predicting

Predictions of the Churn Guard Pro can be shown as follows:

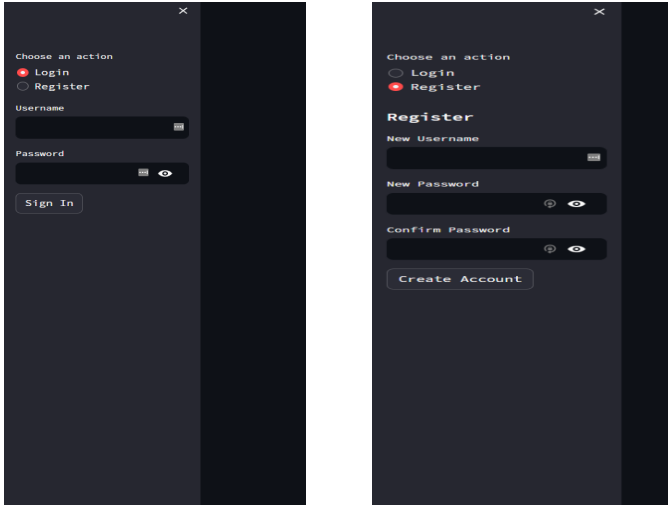


Fig. 16. Login and Register screens

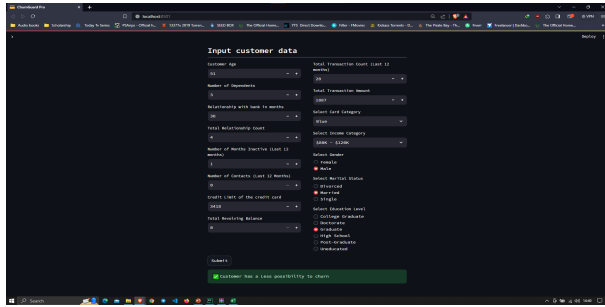


Fig. 17. Web application prediction

C. Testing

- Unit Testing: Unit testing was done on individual functional units of the web application to ensure they perform accordingly.
- Integration Testing: Integration testing was performed to ensure that the functional units of the system work together.
- System Testing: System testing was performed to ensure that the system is performed accordingly and meets the requirements.

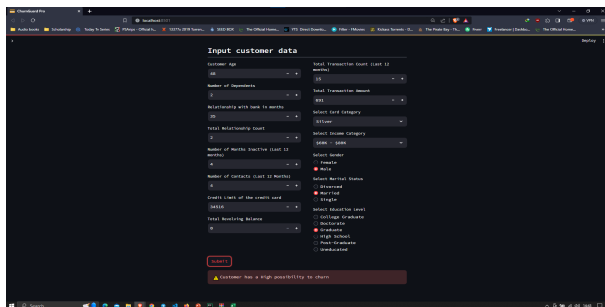


Fig. 18. Web application prediction

VI. CONCLUSION

This research has made significant contributions to the field of customer churn prediction by developing an effective XGBoost machine learning model and deploying it within a user-friendly web application. In terms of predicting customer attrition, the XGBoost model performed remarkably well, with an overall accuracy of 95.7%, a precision of 96.9%, and a recall of 97.9%. These findings demonstrate the model's accuracy in differentiating between consumers who have churned and those who have retained, offering useful information to companies looking to grow their clientele.

The XGBoost model's effectiveness is further increased by creating a web application with the Streamlit framework. A seamless and safe user experience is guaranteed by the web application's dynamic input processing capabilities, user authentication system, and session state management features.

The research's conclusions provide insightful advice for companies looking to put into practice efficient customer churn prediction techniques. Businesses can utilize the user-friendly online application and the better performance of the XGBoost model to identify at-risk consumers and undertake targeted interventions to reduce churn rates. Businesses can increase their overall profitability and client lifetime value by utilizing these technologies.

Subsequent investigations could focus on integrating supplementary data sources, such as customer feedback and social media data, to improve the precision of churn prediction models. Furthermore, investigating the use of deep learning methods for churn prediction in customers may yield fresh perspectives and enhance model performance.

All things considered, this study has improved our knowledge of customer churn prediction and given firms useful tools to deal with this pressing problem. The results have ramifications for many different industries where keeping customers is essential to success, such as banking, telecommunications, retail, and e-commerce.

ACKNOWLEDGMENT

Special acknowledgments go to Dr.Sampath Edirisinghe and all the other academic members of the Department of Computer Engineering, Faculty of Engineering, University of Sri Jayewardanepura, for guiding me throughout this project to success.

REFERENCES

- [1] American Banker, "Customer Churn: The Silent Killer of Your Bank." [Online]. Available: <https://www.americanbanker.com/news/why-businesses-are-growing-frustrated-with-the-service-at-their-bank>
- [2] H. C. Homburg, J. C. Koslow, and S. B. Mukherjee, "The Value of Keeping the Right Customers." *Harvard Business Review*, vol. 92, no. 10, pp. 102-109, 2014.
- [3] Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., and Chatzisavvas, K. Ch. (2015). Predicting credit card customer churn using machine learning techniques. *Knowledge and Information Systems*, 35(2), 539-552.
- [4] Diamantaras, K. I., and Kung, S. Y. (2019). Computer-assisted customer churn management: State-of-the-art and future trends. *Decision Support Systems*, 115, 174-195.