Exploring Statistical Machine Learning Algorithms for Predicting Customer Churn

1st Oindri Kar,
Engineering Systems and Computing,
University of Guelph,
Ontario, Canada
okar@uoguelph.ca

***Abstract*— Customer churn is the term used to describe the phenomenon of customers or subscribers abandoning a business or service. Customers in the telecommunications sector have the ability to actively switch between multiple service providers by making their own choices. To reduce customer turnover, telecom companies must identify the clients who are most likely to leave. Our main objectives are to tackle the most significant obstacles in customer churn prediction, apply sophisticated preprocessing methods to guarantee the quality of the dataset, establish a foundation for precise analysis, and investigate Machine Learning Algorithms with a focus on assessing, analyzing, optimizing, and evaluating the results.**

Keywords— customer churn, machine learning algorithms, SVM, random forest, logistic regression, customer attrition.

I. INTRODUCTION

Customer attrition is one of the biggest problems and top worries for big businesses. Due to the direct effect on business profits, especially in the telecommunication sector, businesses are trying to create techniques to anticipate possible client attrition. Thus, identifying the causes of customer attrition is essential to taking the appropriate steps to lower it. [1].

The main focus of our extensive study is implementing machine learning methods, specifically aiming to tackle the crucial problem of telecom industry - customer churn prediction. We start by addressing the critical challenge of predicting customer churn and recognize the importance of overcoming the obstacles associated with this predictive task. Our study emphasizes the importance of advanced preprocessing techniques in order to conduct a robust and accurate analysis. The exploration of various Statistical Machine Learning Algorithms, such as Support Vector Machines (SVM), Random Forest, Logistic Regression is part of our research methodology [8].

Each algorithm is chosen for its distinct characteristics and potential application to customer churn prediction. The research delves into the intricate details of each algorithm, analyzing, fine-tuning, and assessing its success in predicting the churn. To assess the efficacy of these algorithms, we use confusion metrics, a broad set of evaluation metrics that includes accuracy, precision, recall, and F1-score [5].

We hope that by using visual representations, we can provide a clear and intuitive understanding of the complex relationships and patterns revealed by our analysis.

In essence, our research is more than just an investigation into predictive modeling; it is a strategic effort to provide telecom companies with actionable insights derived from a nuanced understanding of customer churn dynamics [3].

The following is how the paper's structure unfolds:

Section I: Following the Introduction, the Project Motivation and Objectives are described in this section.
Section II: Prior research in the area of customer churn prediction is concisely summarized.
Section III: Outlines the Project Methodology, describing the approach taken to put the code into practice.
Section IV: Comprises the outcomes of several algorithms' implementations.
Section V: Based on the algorithms employed, a comparative study is carried out.
Section VI: Explores future directions and suggests possible lines of future research.
Section VII: The summary of the major results and inferences drawn from the study to wrap up the project report.

II. LITERATURE REVIEW

A Review on Existing Research:

Customer Churn prediction is an important field of business research, with studies using machine learning algorithms to anticipate client attrition. Previous research has emphasized the need of early detection of disappointment and proactive retention methods. To improve model accuracy and efficiency, several datasets and approaches are used [4].

ML Algorithm-Based Research:

Numerous studies employ SVM, Random Forest, Decision Trees, Logistic Regression, and KNN to forecast churn. SVM excels in capturing complicated decision boundaries, Random Forest and Decision Trees give interpretability, Logistic Regression acts as a benchmark, and KNN captures local patterns well [2].

Gaps in Literature:

Even with advancements, there remain gaps:

- Limited Comparative Analysis: Detailed analyses of algorithm performance are necessary.
- Dataset Diversity: Research across many features is necessary due to inconsistent dataset types.

- Interpretability: Improve the interpretability of SVM and neural networks.
- Dynamic Feature relevance: Analyzing the evolution of feature significance.

III. METHODOLOGY

A. *Data Collection*

The dataset under consideration is collected from a Telecommunication company in California in Q2 2022 and comprises a comprehensive set of information, encompassing 7034 entries across 21 distinct columns. These columns encapsulate a variety of aspects relevant to customer interactions and characteristics [14].

Column Descriptions:

- Churn (Target): Indicates whether customers have terminated their services within the last month. This serves as the target variable for analysis.
- Services: Enumerates the diverse services subscribed to by each customer, such as phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- Customer account information includes tenure, contract type, payment method, paperless billing, monthly charges, total charges, etc.
- Demographic Information consists of gender, Age, and partner and dependent information.

B. *Exploratory Data Analysis*

By visualizing patterns and trends, data visualization improves insights in customer churn prediction. Key indications and correlations can be found with the aid of visualizations such as heatmaps and line charts. While bar graphs efficiently compare demographic data, pie charts effectively demonstrate the distribution of services. In general, data visualizations make data easier to interpret intuitively, which helps decision-makers reduce customer attrition.

1. *Heatmap Correlation Matrix*

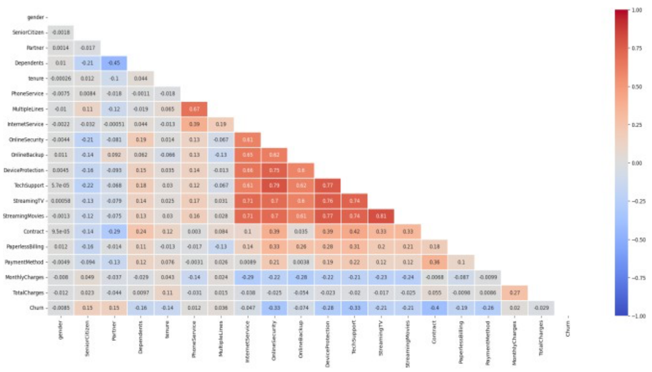

Figure 1: Heatmap Showing Correlations

The correlation matrix for our dataset has been shown. The degree to which two variables change together is described statistically by correlation.

In other words, it measures the degree to which two variables are related. The correlation coefficient is a number that ranges from -1 to 1.:

- +1 indicates a perfect positive correlation (as one variable increases, the other also increases).
- -1 indicates a perfect negative correlation (as one variable increases, the other decreases).
- 0 indicates no correlation.

The correlation analysis with "Churn" reveals notable associations. Key factors positively correlated with churn include monthly charges, paperless billing, and senior citizenship. Conversely, features negatively correlated with churn include contract length, tenure, and the availability of tech support and online security services. These insights suggest that customers with longer-term commitments and access to support services are less likely to churn.

2. *Distribution of Churn and Gender*

Figure 2 Gender and Churn Distribution

Analysis from plots: 26.6 % of customers switched to another firm. 49.5 % customers are female and 50.5 % are male.

3. *Proportion of Senior Citizens*

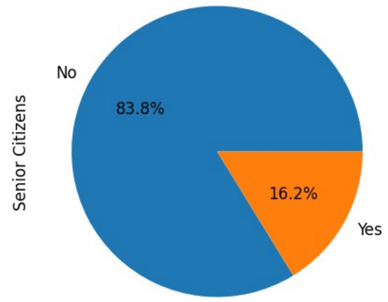

Figure 3: Percentage of Senior Citizens

There are only 16% of the customers who are senior citizens. Thus, most of our customers in the data are younger people.

4. *Number of Customer by Contract Type*

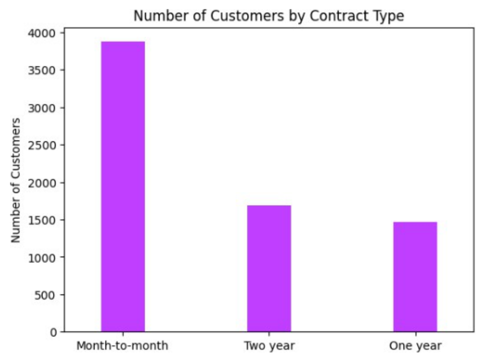

Figure 4: Number of Customer by contract type

The analysis of customer contracts reveals a noteworthy distribution, with a significant majority opting for a month-to-month contract. Specifically, the data indicates that a substantial portion of our customers (most of them) prefer the flexibility of a month-to-month commitment. Interestingly, the enrollment numbers for both one-year and two-year contracts are comparable, suggesting a balanced preference among customers for longer-term commitments, perhaps driven by specific service or pricing considerations.

5. *Churn by Contract Type*

The breakdown of customer churn by contract type unveils distinctive patterns. Notably, the month-to-month contract exhibits a higher churn rate at 43%, emphasizing the susceptibility of this group to discontinuation. In contrast, the one-year contract witnesses an 11% churn, while the two-year contract experiences a minimal 3% churn. Impressively, the remaining percentages denote no churn, underscoring the stability and loyalty associated with longer-term contractual commitments.

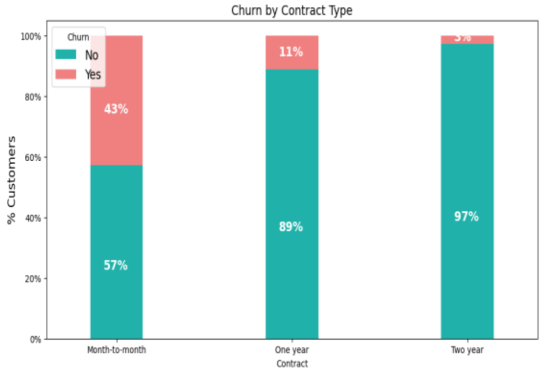

Figure 5: Churn by contract type

6. *Payment Method Distribution*

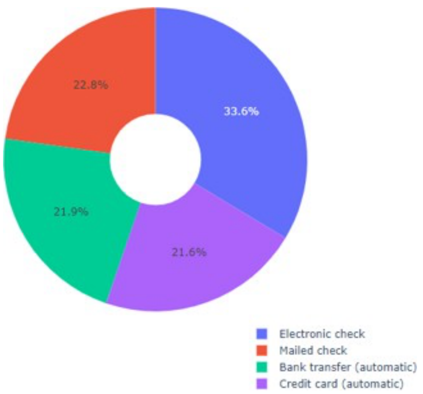

Figure 6: Payment Method Distribution

The distribution of payment methods among our customers reveals a diversified landscape. A notable percentage of 33.6% opt for the convenience of credit card payments, demonstrating a prevalent preference. Additionally, electronic cheques are utilized by 21.6%, while mailed cheques account for 22.8%. Bank transfers also play a significant role, constituting 21.9% of the overall payment method distribution. This variety in payment preferences shows the importance of offering multiple options to cater to the diverse needs of our customer base.

7. *Payment Method Distribution w.r.t. churn*

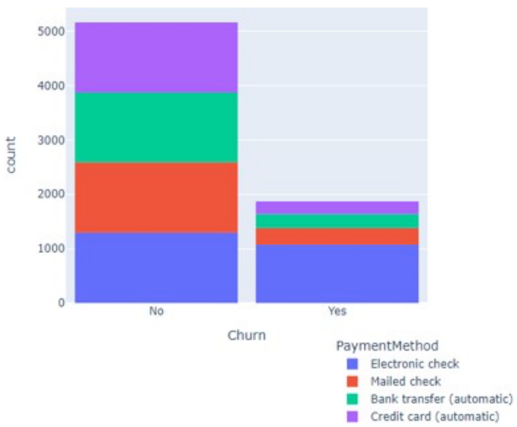

Figure 7: Payment Method Distribution w.r.t. churn

The analysis of the churn by payment method exposes notable disparities. Customers using electronic cheques exhibit the highest churn, signaling potential friction or dissatisfaction with this payment approaches. In contrast, those utilizing credit cards display the lowest churn, highlighting the stability associated with this payment method.

8. *Distribution of total charges by churn*

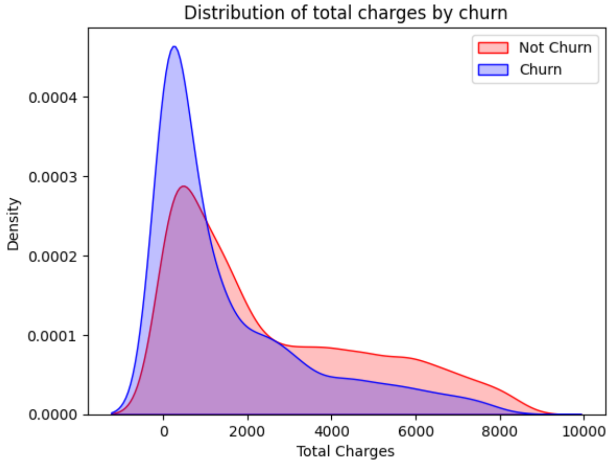

Figure 8: Distribution of total charges by churn

Elevated churn among the customers with lower total charges is likely influenced by their price sensitivity and enrollment in less comprehensive plans with limited features. This segment may be inclined to explore alternatives or upgrade for enhanced services. Implementing targeted strategies, including personalized promotions and feature-rich offerings, can effectively address these dynamics and bolster customer retention.

9. *Distribution of Monthly charges by churn*

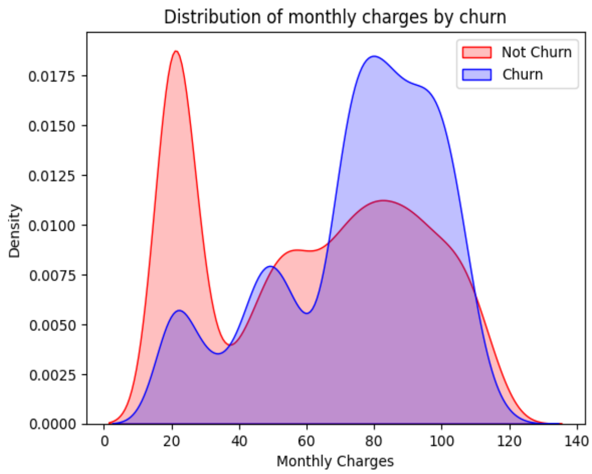

Figure 9: Distribution of Monthly charges by churn

The distribution of monthly charges unveils a noteworthy trend, indicating a positive correlation between higher monthly charges and increased churn. Customers with elevated monthly fees appear more susceptible to discontinuation. This relationship underscores the importance of balancing pricing strategies to mitigate churn risks, with considerations for personalized value propositions or more competitive offerings for high-paying customers.

10. Churn Distributions for Different Features

In the churn distribution graphs for gender, partners, senior citizens, and phone service, we discern insightful patterns in customer behavior. Notably, these visualizations provide a clear overview of how different demographic and service-related factors influence churn. These insights pave the way for targeted retention strategies tailored to diverse customer segments."

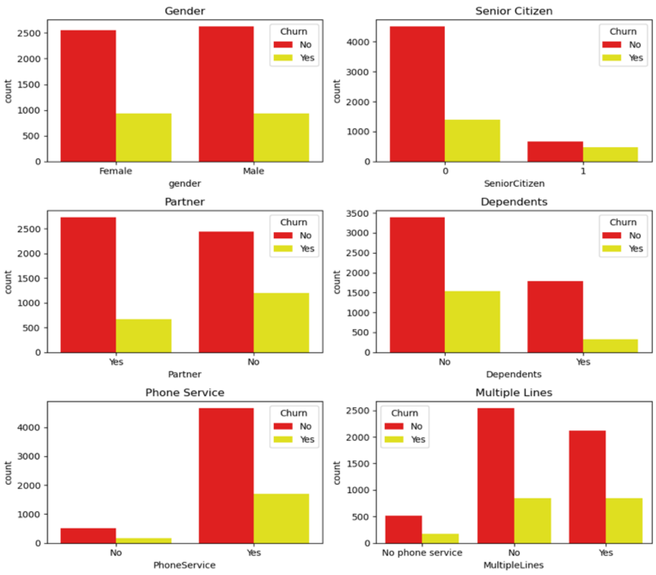

Figure10: Churn for different features

C. Data Preprocessing

Preparing the data is essential to cleaning up unprocessed datasets. The steps that follow outline a methodical strategy to improve data quality and streamline subsequent analysis procedures. [10].

1. Handling Missing Data: It is essential to find and fix any missing data. To guarantee data completeness, strong techniques are used, such as deleting rows with values that are missing or imputing them using statistically reliable measures like the mean. 11 null values in the Total Charges column were changed to the mean of the values.

2. Label Encoding: To include categorical variables in analytical models, label encoding is applied judiciously. By methodically converting categorical data into a numerical representation, this procedure makes it possible for machine learning algorithms to integrate with simplicity.

3. Standardization: Achieving homogeneity in data ranges requires standardizing numerical features. By utilizing standard scalar approaches, numerical attributes are standardized to reduce the effect of different magnitudes and promote fair participation in further investigations.

To summarize, the thorough preparation stages help to improve datasets and establish the foundation for accurate and dependable analysis.

D. Feature Selection and Engineering

Feature Selection:

The column "Customer_id" was dropped because it had no significant contribution to the prediction of the churn variable [7].

Feature Engineering:

Categorical columns for one-hot encoding were identified, encompassing 'PaymentMethod,' 'Contract,' and 'InternetService.' Simultaneously, numerical columns were enumerated for subsequent standardization using StandardScaler, while residual categorical columns underwent label encoding. This structured approach, orchestrated through a ColumnTransformer, guarantees that various data formats are preprocessed consistently, which promotes the development of robust machine learning models. [5].

E. Experimental Setup:

Train-Test Split and Cross-Validation Strategy (70, 15, 15):

The dataset was divided into training, validation, and test sets using a stratified train-test split with a ratio of 70% training data, 15% validation data, and 15% test data. Stratification ensures that the distribution of the target variable (churn status) is preserved across the sets, preventing biases in the model evaluation process.

During the model training phase, a cross-validation method was used to further improve the resilience of the model and reduce overfitting. To be more precise, k-fold cross-validation was applied with k=5. To do this, split the training set into five subsets, use four out of the five subsets of data to train the model, and use the remaining five to validate the model. This process is performed five times, and the model's performance is averaged across all folds. [9].

Evaluation Metrics Chosen:

The evaluation of the models was conducted using a comprehensive set of metrics to provide a nuanced understanding of performance [4].

- Accuracy: It measures the overall correctness of the model predictions and is suitable for balanced datasets.
- Precision: It assesses the accuracy of positive instances among all positive predictions, emphasizing the reliability of the model when it predicts positive instances.
- Recall (Sensitivity): It quantifies the ability of the model to correctly identify all relevant instances, particularly relevant in scenarios where false negatives are costly.
- F1-Score: The harmonic means of precision and recall, providing a balanced measure of a model's precision and recall. Especially when the classes are imbalanced.
- Specificity (True Negative Rate): It measures the model's proficiency in identifying negative instances accurately and is pertinent in situations where false positives are critical.

The selection of these metrics ensures a comprehensive assessment of the model's performance across various aspects, catering to both the model's predictive accuracy and its ability to handle class imbalances [6].

F. *Implemented Machine Learning Algorithms*

We have implemented a total of five algorithms on the customer churn prediction dataset as follows:

i) SVM (Support Vector Machine)

Support Vector Machine (SVM) is a powerful observational machine learning algorithm used for regression and classification tasks. It determines the best hyperplane for classifying the data points to minimize the distance and margin between classes. SVMs can use tricks to map the data into a dimensional space, which makes it suitable for non-distributive data. SVM can achieve high discriminative power by creating a multidimensional hyperplane that clearly separates two classes, which can be modified according to different datasets and improve accuracy [5].

Support Vector Machines (SVM) are valuable in customer churn prediction due to their ability to effectively handle complex, non-linear relationships within data. SVMs optimize the margin between different classes, allowing them to discern intricate patterns and capture nuanced dependencies present in churn-related features. Additionally, SVMs inherently handle outliers well, contributing to a robust model performance. Their effectiveness in capturing complex decision boundaries and handling imbalanced datasets further enhances their utility in predicting customer churn scenarios [12].

ii) Logistic Regression

Despite its name, logistic regression is a classification algorithm used for two-class and multi-class classification problems. It predicts the probability that a sample belongs to a certain class. A logistic function (sigmoid function) is used to map the predicted values from 0 to 1. Logistic regression assumes a linear relationship between the characteristics of the exponential variable and the log-likelihood.

Logistic regression is useful for predicting customer attrition since it is simple and easy to interpret. It offers a clear comprehension of the connection between the likelihood of churn and independent variables. The binary nature of the outcome (churn or not) aligns with the typical nature of churn prediction tasks. Logistic regression is robust with smaller datasets, commonly encountered in churn analysis. Its coefficients offer insights into the impact of each predictor on the likelihood of churn, aiding in the identification of significant features. Furthermore, logistic regression facilitates easy integration into business decision-making processes, making it a pragmatic choice for customer churn modeling [11].

iii) Random Forest

The Random Forest algorithm works as an ensemble learning method using many decision trees. The final forecast or regression result is determined by the class mean of individual tree forecasts. Combining multiple decision tree products provides greater robustness and accuracy. Random Forest can solve overfitting concerns and have the capability to handle categorical data and numeric data [8].

Random Forest is highly advantageous for customer churn prediction due to its ensemble learning approach, combining multiple decision trees for robust predictions. It excels in capturing complex relationships and interactions within diverse datasets, offering improved accuracy compared to individual models. The inherent ability to handle both numerical and categorical features makes it versatile for customer data, where varied factors influence churn. Random Forest mitigates overfitting by aggregating predictions, enhancing generalization to unseen data. Feature importance estimation aids in identifying critical factors influencing churn, aiding strategic decision-making. Furthermore, its scalability accommodates large datasets, ensuring efficiency in handling the voluminous customer data often encountered in predictive analytics for churn [8].

Feature Importance:

A feature importance graph visually communicates the contribution of each feature to a machine learning model's predictive performance. This tool aids in pinpointing features with the most significant impact on predictions, guiding attention to key determinants. It enhances interpretability by quantifying the influence of individual features, facilitating a nuanced understanding of their importance. This insightful analysis is valuable for prioritizing efforts and optimizing models by focusing on the most influential factors. Ultimately, feature importance graphs offer a concise and accessible means to unravel the complexity of a model's feature dynamics [6].

The most important features are customer tenure, total charges, monthly charges, and contract type. The least important feature includes gender, partner, dependencies, phone services, etc.

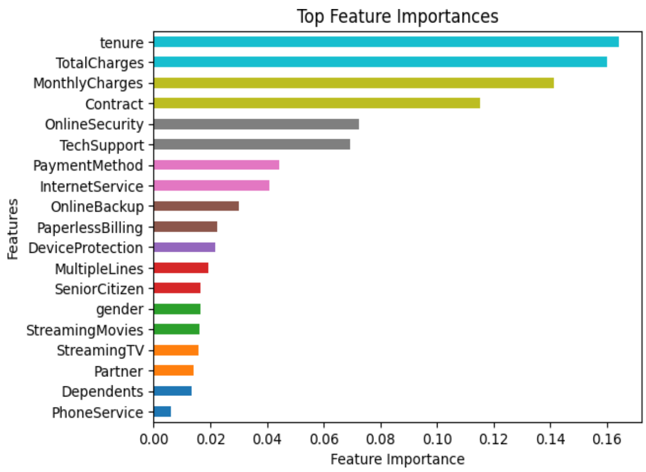

Figure 10: Feature Importance graph

G. *Hyper-Parameter Tuning for Each Algorithm using Grid Search CV:*

In pursuit of optimizing model performance, hyper-parameter tuning was systematically conducted for each machine learning algorithm, employing Grid Search Cross-Validation (CV) methodology. Searching through all the possible combinations in order to find the best set of hyperparameters according to evaluation metric in this case accuracy, F1 Score, precision and specificity.

i) SVM Parameter Tuning:

Best SVM Parameters: {'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}. The SVM model underwent rigorous parameter tuning. The choice of the kernel, regularization parameter (C), and kernel-specific parameters like 'gamma' were systematically explored. The optimal configuration, {'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}, was identified through Grid Search CV, ensuring robustness in capturing non-linear relationships within the data.

i) RF Parameter Tuning:

Best Parameters: {'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_depth': 10}

The Random Forest model was fine-tuned by varying parameters such as the number of trees ('n_estimators'), maximum depth ('max_depth'), and minimum samples per leaf ('min_samples_leaf' and 'min_samples_split'). The resulting optimal configuration reflects a balance between model complexity and predictive accuracy.

ii) Logistic Regression Parameter Tuning:

Best Logistic Regression Parameters: {'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}

Logistic Regression parameters, notably regularization strength ('C'), penalty term ('penalty'), and solver type ('solver'), were systematically adjusted to achieve optimal model performance. The resulting configuration attests to the model's ability to balance bias and variance effectively.

This comprehensive hyper-parameter tuning approach, anchored in the principles of Grid Search CV, establishes refined configurations for each algorithm, enhancing their predictive capabilities and generalizability.

IV. RESULTS

Implemented Algorithm	Testing Accuracy	Validation Accuracy	Precis ion	Recall	F1-Score
SVM	79.28	79.45	0.63	0.53	0.58
Logistic Regression	79.19	79.26	0.62	0.56	0.59
Random Forest	77.96	80.68	0.60	0.50	0.55

Table 1: Evaluation metrics for each model

The above evaluation metrics show that the Random Forest Algorithm performs the best followed by Logistic Regression SVM (Statistical Vector Machine) having the lowest performance on predicting the possibility of customer churn.

It is also the desired result as Random Forest is an Ensembled Algorithm, as it a combination of many Decision trees and then the best result is taken as the result. The other models, Logistic Regression and SVM perform well when the data is linearly separable. In this case the dataset used contains many attributes, where not all of them are linearly separable.

V. COMPARATIVE ANALYSIS

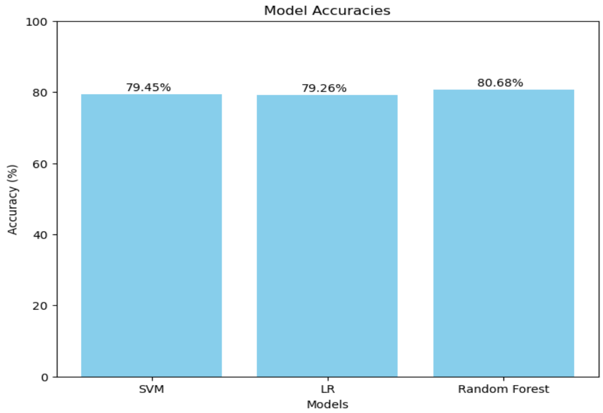

Figure 11: Validation set accuracies.

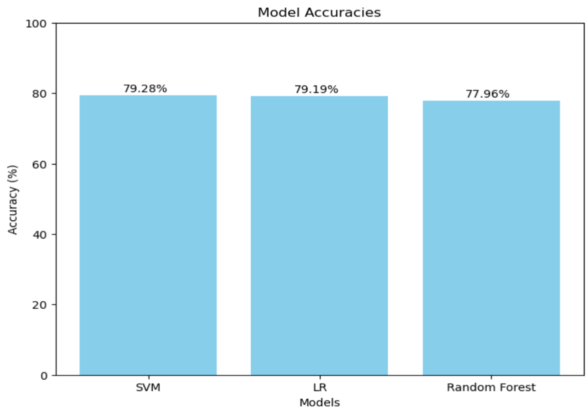

Figure 12: Test set accuracies

The above comparative graphs show that the test and validation set accuracies. The best-performing model is Random Forest model which can be concluded from the fact that the model performs similarly on both the validation and test sets. This is often the desired outcome, as it suggests that the model generalizes well to new, unseen data.

VI. FUTURE SCOPE

The prediction system has allowed to come up with some success strategies that can help reduce customer churn. These are explained below [9]:

- Optimize Contract Structures: Encourage customers to opt for longer-term contracts through incentives and discounts and provide flexibility in contract terms to cater to diverse customer preferences.
- Competitive Pricing and Value: Regularly assess and adjust pricing to remain competitive and introduce flexible plans and promotional offers to enhance customer value.
- Enhance Customer Satisfaction: Conduct regular customer satisfaction surveys to identify pain points and address issues related to customer service, network quality, and overall user experience.
- Personalized Engagement: Analyze usage patterns of customers who churned and provide personalized recommendations and offer promotions tailored to individual needs and preferences.
- Retention Programs for New Customers: Implement programs targeting new customers to increase loyalty and provide incentives and rewards for customers with shorter tenure.

- Promote Additional Services: Highlight the value of additional services such as tech support and device protection and consider bundling services for a comprehensive offering.
- Convenient Payment Options: Offer various payment methods to cater to customer preferences and incentivize the use of certain payment methods for ease of transaction.
- Long-Term Contracts with Autopay: Encourage customers to set up automatic payments for convenience and provide discounts for customers opting for longer-term contracts with auto-pay.
- Tailored Communication for Senior Citizens: Customize communication and services to suit the preferences and requirements of senior citizens and consider special promotions and offers for this demographic.
- Proactive Communication Strategies: Implement targeted communication strategies such as personalized newsletters and proactive outreach and address potential concerns before they escalate.
- The future work of this system included using k-fold cross-validation and using the validation set [11].

VII. CONCLUSION

In this project, we explored diverse algorithms - Logistic Regression, Random Forest, and SVM. The importance of a comprehensive strategy in addressing customer churn is emphasized. Actionable insights for businesses to tailor effective retention strategies are provided. The importance of this type of research in the telecom market is to assist companies in increasing their profits. Predicting the churn has become one of the most important sources of revenue for telecom companies [1]. Using the modern machine learning algorithms for statistical analysis will be better. It helps the company know their customers better and work towards satisfying their needs. The main reason behind customers leaving a service provider and joining a new one, is the availability of better offers. If the company can predict the customer requirements in advance, then they will be able to retain them. According to studies it shows that the cost of acquiring a new customer is 25% higher than retain an existing customer. So, if the telecom companies use the predictive analysis they will be able to lower their revenue costs and spend the saved money on development.

Lastly, we acknowledge the ongoing significance of machine learning in customer relationship management. The Machine Learning Models can be used for predictive analysis to know which factors affect customer attrition rates. In the modern competitive telecommunication business, it’s very important to know factors that could lead to losing a customer and working towards retaining them. It a highly competitive market and focusing on promising factors can help the company strive in the business.

REFERENCES

1. S. Upadhyay and R. M, "Customer Churn Prediction using Machine Learning," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023,pp.17,doi:10.1109/ICCCNT56998.2023.10307748
2. F. Alhaqui, M. Elkhechafi and A. Elkhadimi, "Machine learning for telecoms: From churn prediction to customer relationship management," 2022 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), Soyapango, El Salvador, 2022, pp. 1-5, doi: 10.1109/ICMLANT56191.2022.9996496.
3. Raj and D. Vetrithangam, "Machine Learning and Deep Learning technique used in Customer Churn Prediction: - A Review," 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2023, pp. 139-144, doi: 10.1109/CISES58720.2023.10183530.
4. N. I. A. Razak and M. H. Wahid, "Telecommunication Customers Churn Prediction using Machine Learning," 2021 IEEE 15th Malaysia International Conference on Communication (MICC), Malaysia, 2021,pp.81-85,doi: 10.1109/MICC53484.2021.9642137.
5. Kaur and J. Kaur, "Customer Churn Analysis and Prediction in Banking Industry using Machine Learning," 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, 2020, pp. 434-437, doi: 10.1109/PDGC50313.2020.9315761.
6. L. Butgereit, "Work Towards Using Micro-services to Build a Data Pipeline for Machine Learning Applications: A Case Study in Predicting Customer Churn," 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE), Aswan, Egypt, 2020, pp. 87-91, doi: 10.1109/ITCE48509.2020.9047807.
7. Z. Chen and Z. Fan, "Building comprehensible customer churn prediction models: A multiple kernel support vector machines approach," ICSSSM11, Tianjin, China, 2011, pp. 1-4, doi: 10.1109/ICSSSM.2011.5959439.
8. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," in IEEE Access, vol. 7, pp. 60134-60149, 2019, doi: 10.1109/ACCESS.2019.2914999.
9. P. Gopal and N. B. MohdNawi, "A Survey on Customer Churn Prediction using Machine Learning and data mining Techniques in E-commerce," 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Brisbane, Australia, 2021, pp. 1-8, doi: 10.1109/CSDE53843.2021.9718460.

10. V. Chang, X. Gao, K. Hall and E. Uchenna, "Machine Learning Techniques for Predicting Customer Churn in A Credit Card Company," 2022 International Conference on Industrial IoT, Big Data and Supply Chain (IIoTBDSC), Beijing, China, 2022, pp. 199- 207, doi: 10.1109/IIoTBDSC57192.2022.00045.
11. M. H. Seid and M. M. Woldeyohannis, "Customer Churn Prediction Using Machine Learning: Commercial Bank of Ethiopia," 2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), Bahir Dar, Ethiopia, 2022, pp. 1-6, doi: 10.1109/ICT4DA56482.2022.9971224.
12. Ş. Dönmez, "Machine Learning-Based Merchant Churn Prediction in Banking," 2023 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Turkiye, 2023, pp. 503-508, doi: 10.1109/UBMK59864.2023.10286753.
13. M. A. Hassonah, A. Rodan, A. -K. Al-Tamimi and J. Alsakran, "Churn Prediction: A Comparative Study Using KNN and Decision Trees," 2019 Sixth HCT Information Technology Trends (ITT), Ras Al Khaimah, United Arab Emirates, 2019, pp. 182-186, doi: 10.1109/ITT48889.2019.9075077.
14. Dataset Link: https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113
15. J Jain, H., Khunteta, A., Srivastava, S. (2021). Telecom churn prediction and used techniques, datasets and performance measures: a review. Telecommunication Systems, 76(4), 613-630.
16. Vo, N. N., Liu, S., Li, X., Xu, G. (2021). Leveraging unstructured call log data for customer churn prediction. Knowledge-Based Systems, 212, 106586 .
17. De Caigny, A., Coussement, K., Verbeke, W., Idbenjra, K., Phan, M. (2021). Uplift modeling and its implications for B2B customer churn prediction: A segmentation-based modeling approach. Industrial Marketing Management, 99, 28-39.
18. V.S.Y.Lo,Thetrueliftmodel:anoveldataminingapproachtor esponse modeling in database marketing, SIGKDD Explor. Newsl. 4 (2002) 78–86
19. LU, Junxiang et PARK, O. Modeling customer lifetime value using survival analysis—an application in the telecommunications industry. Data Mining Techniques, 2003, p. 120-128.
20. Turban,Aronson,Liang,andSharda,2007;Bersonetal.,2000 ;Lejeune ,2001; Ahmed, 2004 and Berry and Linoff ,2004; Lau, Wong, Hui, amp; Pun, 2003.