**PREDICTIVE MODELING OF CUSTOMER CHURN IN THE TELECOMMUNICATIONS INDUSTRY: LEVERAGING MACHINE LEARNING AND PREPROCESSING TECHNIQUES.**

**Author’s Name: DAIVIN NGANGA NJOROGE**

**Date: 29/02/2024**

**Contact Info:** [**daivinnganga88@gmail.com**](mailto:daivinnganga88@gmail.com)

**LinkedIn:** [**https://www.linkedin.com/in/daivin-nganga-247512261/**](https://www.linkedin.com/in/daivin-nganga-247512261/)

**GitHub:** [**https://github.com/Statwriter22**](https://github.com/Statwriter22)

**Abstract**

This research project investigates the application of machine learning (ML) algorithms in predicting customer churn within the telecommunications industry.

The study evaluates the performance of six ML algorithms, including Random Forest, Logistic Regression, Support Vector Machine, Decision Trees, KNeighbors, and Neural Networks, using a comprehensive dataset comprising customer attributes, service usage metrics, and churn status.

Preprocessing techniques such as label encoding, one-hot encoding, and standard scaling are applied to prepare the data for modeling. Visualizations, including Kernel Density Estimation (KDE) plots and correlation matrices, are utilized to analyze the distribution of features and relationships between variables.

The results reveal variations in prediction accuracy among the ML algorithms, with Support Vector Machine achieving the highest accuracy of 80%.

Additionally, strong correlations are observed between tenure and total charges, moderate correlations between total charges and monthly charges, and weaker associations between monthly charges and tenure. These findings provide valuable insights into churn behavior and underscore the importance of incorporating both traditional predictive features and advanced analytical techniques in developing robust churn prediction models within the telecommunications industry.

**Research Question**

How do different machine learning algorithms perform in predicting customer churn within the telecommunications industry, and to what extent can preprocessing techniques and visualization assist in comprehending churn behavior and determining the optimal deployment model?

**INTRODUCTION**

Customer Churn also known as the rate of attrition is defined as the number of customers or subscribers that stopped using your company’s product or service during a given period of time (Frankenfield, 2022). It is one of the most important business metrics for a business because it is much less expensive to retain existing customers than it is to acquire new customers

In the field of telecoms, we can refer Churn rate as the percentage of customers who close their contracts or subscriptions with your company in any given time period (Frankenfield, 2022). Customers in this field can choose from a variety of service providers and actively switch from one to the next.

A high churn rate means you aren’t able to retain your customers implying a low loyalty to the company services. In a positive view, it is a good indicator there is a flaw in the services offered and customers are looking elsewhere.

To reduce churn rate, telecom companies need to predict which customers are at a high risk of churn. By addressing churn these businesses may not only preserve their market position, but also grow and thrive.

**Main Objective.**

To evaluate the performance of various machine learning algorithms in predicting customer churn within the telecommunications industry and assess the effectiveness of preprocessing techniques and visualization in understanding churn behavior and selecting the best deployment model.

**Specific Objectives.**

1. Compare the accuracy of different machine learning algorithms, including Random Forest, Support Vector Machine, Logistic Regression, Decision Tree, K-Nearest Neighbors, and Neural Networks, in predicting customer churn.
2. Investigate the impact of preprocessing techniques such as label encoding, one-hot encoding, and standard scaling on the predictive performance of machine learning models.
3. Utilize visualization methods to analyze the distribution of features, correlation between variables, and churn rates in the dataset.
4. Identify the best-performing machine learning algorithm and deployment model based on evaluation metrics.
5. Determine the role of preprocessing techniques and visualization in enhancing the understanding of churn behavior and informing strategic decision-making for customer retention efforts within the telecommunications industry.

**DATA DESCRIPTION.**

The telecommunications sector is grappling with the migration of customers from traditional landline services to cable competitors. To address this issue, there is a need to comprehensively grasp the departing customer demographic and the factors driving their decisions. This initiative centers on developing a machine learning model that not only forecasts customer churn but also provides insights into the influencing elements. Additionally, the project aims to propose potential solutions to prevent customer attrition.

A fundamental step involves gaining insights into diverse customer attributes. This entails an examination of the types and pricing of services offered, along with other relationship-oriented factors like age, gender, and senior citizen.

**3.1 Data Collection.**

The source data used in customer churn was obtained from Kaggle. The datasets are divided into the following customer demographics and service subscriptions, financial aspects and customer engagements. In the dataset there are a total of 7045 records and 20 features. I will sample the data and use the first 100 records for convenience.

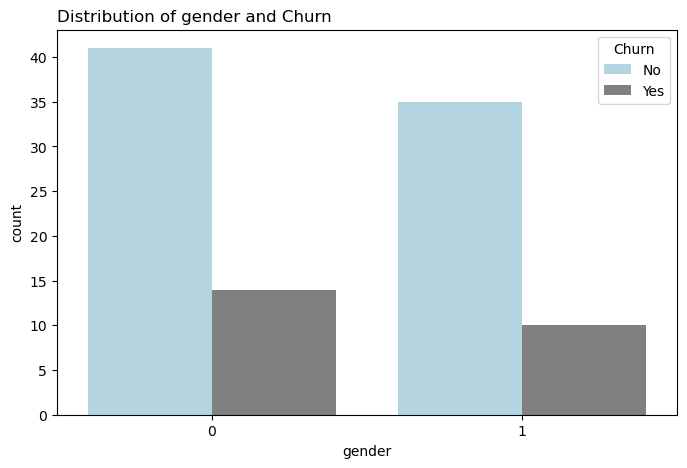
Following are the features present in the dataset:

* customerID - Customer ID
* Gender - Whether the customer is a male or a female
* SeniorCitizen - Whether the customer is a senior citizen or not (1, 0)
* Partner - Whether the customer has a partner or not (Yes, No)
* Dependents - Whether the customer has dependents or not (Yes, No)
* Tenure - Number of months the customer has stayed with the company
* PhoneService - Whether the customer has a phone service or not (Yes, No)
* MultipleLines - Whether the customer has multiple lines or not (Yes, No, No phone service)
* InternetService - Customer’s internet service provider (DSL, Fiber optic, No)
* OnlineSecurity - Whether the customer has online security or not (Yes, No, No internet service)
* OnlineBackup - Whether the customer has an online backup or not (Yes, No, No internet service)
* DeviceProtection - Whether the customer has device protection or not (Yes, No, No internet service)
* TechSupport - Whether the customer has tech support or not (Yes, No, No internet service)
* StreamingTV - Whether the customer has streaming TV or not (Yes, No, No internet service)
* StreamingMovies - Whether the customer has streaming movies or not (Yes, No, No internet service)
* Contract - The contract term of the customer (Month-to-month, One year, Two years)
* PaperlessBilling - Whether the customer has paperless billing or not (Yes, No)
* PaymentMethod - The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
* MonthlyCharges - The amount charged to the customer monthly
* TotalCharges - The total amount charged to the customer
* Churn - Whether the customer churned or not (Yes or No)

**DATA ANALYSIS**

**2.2 Factors influencing customer churn.**

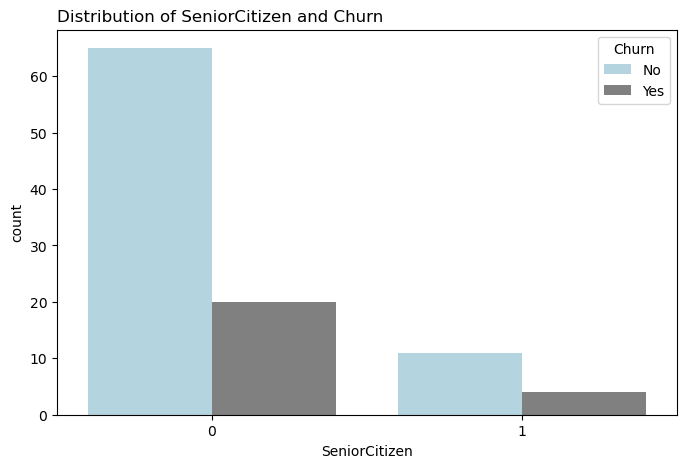
**Gender**

****

The bar chart illustrates a comparison of churn rates between male and female customers in the telecommunications industry. The height of the bar corresponding to females (0) indicates a churn rate of 13, while the bar representing males (1) shows a churn rate of 9.

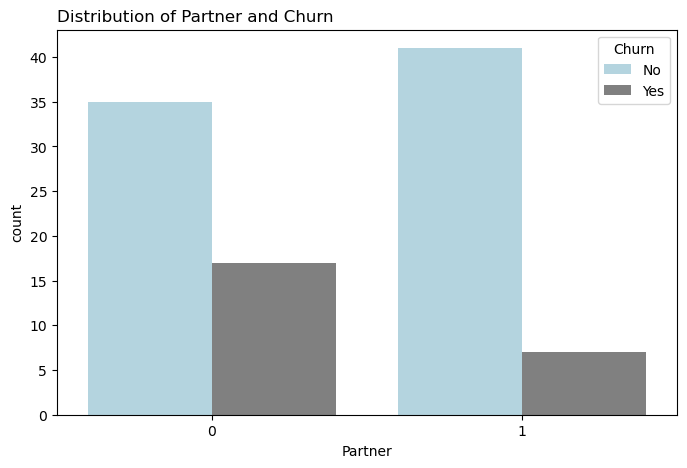
This observation suggests a potential correlation between gender and churn behavior. The higher churn rate among female customers compared to male customers may indicate that gender plays a role in influencing customer attrition. Further analysis is warranted to explore the underlying factors contributing to this difference in churn rates.

**Senior Citizen.**



The bar chart depicting the distribution of senior citizen status reveals that the majority of customers in the dataset are non-senior citizens (0). Interestingly, non-senior citizens exhibit a higher churn rate compared to senior citizens. Despite constituting a smaller portion of the customer base, senior citizens (1) demonstrate a lower churn rate, indicating higher customer loyalty within this demographic.

**Partner.**

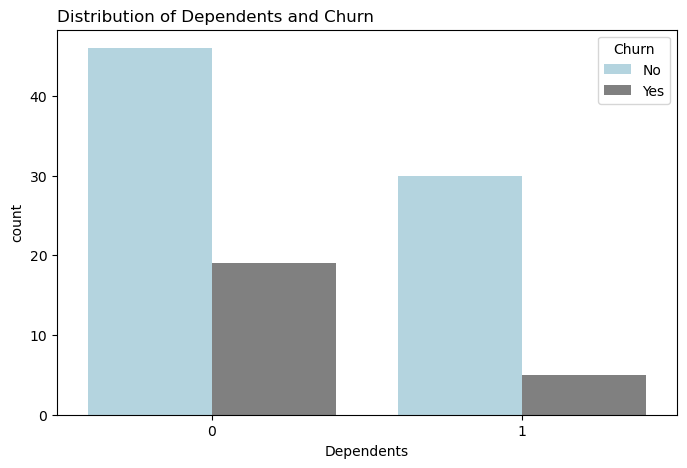
****

The bar chart illustrates that while the majority of customers have partners (1), the churn rate is higher among those who do not have partners (0). Customers who have partners may have stronger ties to the service provider due to shared decision-making or household considerations. As a result, they may be less likely to churn compared to single customers who do not have partners and may be more inclined to switch providers for various reasons.

Having a partner may provide emotional or financial support, which could influence customer satisfaction and loyalty to the service provider. Customers with partners may feel more secure in their decision to stay with the current provider, especially if they have joint accounts or bundled services. Customers without partners may have different usage patterns or needs compared to those with partners. They may be more transient or less satisfied with the services offered, leading to a higher propensity to churn.

Customers without partners may also feel less tied down to a specific provider and may be more willing to explore other options or switch providers if they perceive better value elsewhere. This lack of dependency on the current provider could contribute to the higher churn rate observed among this group.

**Dependents**

****

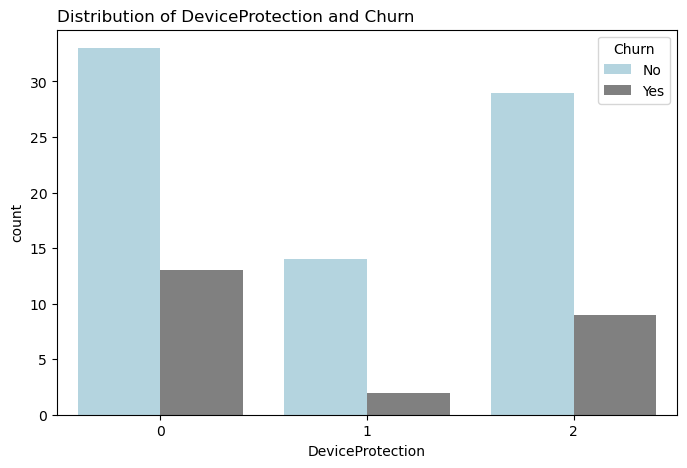
In the analysis of the distribution of dependents and their impact on churn rates, a notable observation emerges: while most customers in the dataset do not have dependents (0), this group exhibits a higher churn rate compared to customers with dependents (1).

This distribution can be interpreted in several ways. Customers without dependents may be more financially independent and flexible in their decision-making. As a result, they may be more inclined to switch providers if they find better deals or perceive greater value elsewhere, leading to a higher churn rate.

Customers with dependents, such as children or elderly relatives, may prioritize stability and continuity in their services to support their family's needs. They may be less likely to churn due to the potential disruption it could cause to their family's routines or services. Having dependents may foster a sense of commitment to the current service provider, as customers may prioritize continuity and reliability in their services to support their family members. This commitment could translate into lower churn rates among customers with dependents.

Customers with dependents may have specific service needs or requirements, such as family plans or bundled services, which could enhance their satisfaction and loyalty to the provider. Conversely, customers without dependents may have more varied needs and preferences, making them more susceptible to churn if their needs are not met.

**Device Protection**

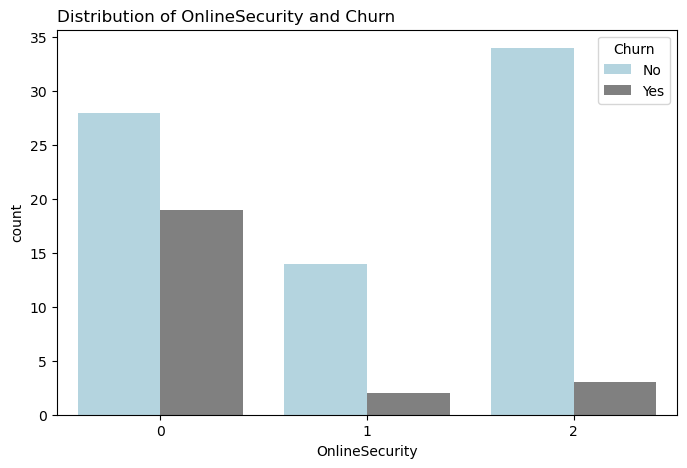
****

The analysis of the distribution of device protection reveals three distinct categories: customers without device protection (0), customers with device protection (1), and customers with no internet service (2). Among these categories, the majority of customers do not have device protection, closely followed by those with no internet service, while a smaller proportion have device protection.

Furthermore, the analysis indicates differences in churn rates across these categories. Despite the relatively low churn rates observed overall, customers without device protection exhibit the highest churn rate, followed by customers with no internet service, while those with device protection have the lowest churn rate.

Customers without device protection or internet service may perceive these offerings as essential for their connectivity or device security. Thus, dissatisfaction or disruptions in these services could prompt them to seek alternative providers, resulting in higher churn rates. Customers with device protection may experience fewer technical issues or concerns, leading to higher satisfaction levels and a lower propensity to churn. Conversely, customers without device protection or internet service may encounter more challenges or frustrations, impacting their loyalty to the provider.

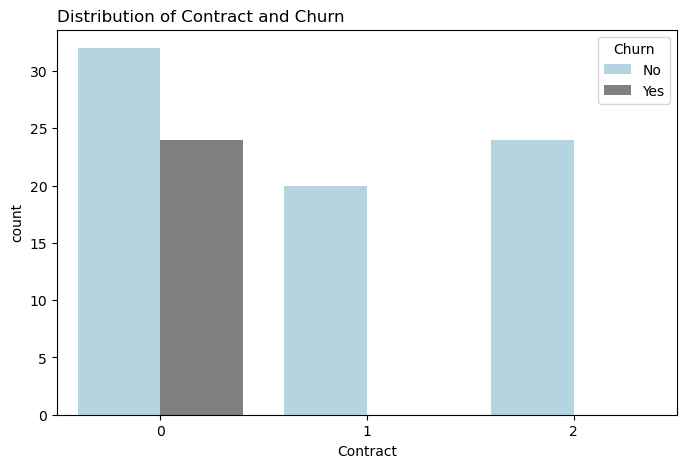
**Online Security**



The analysis of the distribution of online security reveals three distinct categories: customers without online security (0), customers with online security (1), and customers with no internet service (2). Among these categories, the majority of customers do not have internet service, followed by those without online security, while a smaller proportion have online security.

Additionally, the analysis indicates significant differences in churn rates across these categories. Customers without online security exhibit the highest churn rate, representing more than half of the churn instances observed. In contrast, customers with no internet service demonstrate a very low churn rate, while those with online security have an even lower churn rate.

**Contract**

****

The analysis of the distribution of contracts on churn reveals three distinct categories: month-to-month contracts (0), one-year contracts (1), and two-year contracts (2). Among these categories, the majority of customers have month-to-month contracts, followed by two-year contracts, while one-year contracts represent the smallest proportion.

Additionally, the analysis indicates significant differences in churn rates across these contract categories. Customers with month-to-month contracts exhibit the highest churn rate, while the churn rates for one-year and two-year contracts are almost negligible.

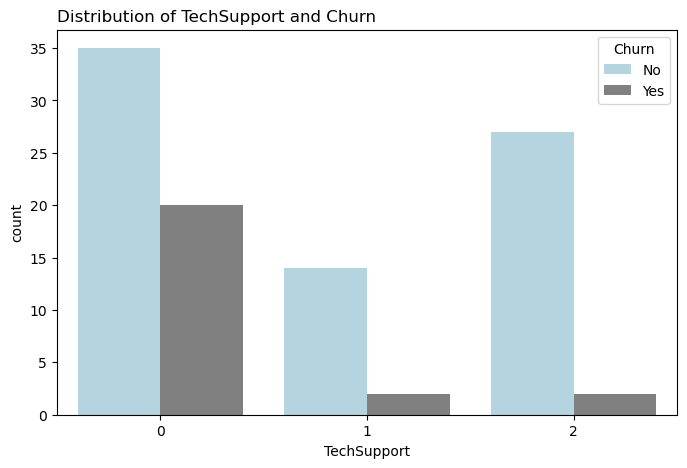
Month-to-month contracts offer customers greater flexibility, allowing them to terminate their service with relatively short notice. As a result, customers with month-to-month contracts may be more prone to churn, as they have fewer barriers to switching providers compared to those locked into longer-term contracts.

Customers with longer-term contracts, such as one-year or two-year contracts, may perceive greater value in their service due to the commitment involved. They may be more likely to remain loyal to the provider throughout the contract duration, as they have made a longer-term commitment and may be less inclined to explore alternative options.

Providers may offer incentives or discounts to encourage customers to sign longer-term contracts, such as reduced monthly fees or promotional offers. These retention strategies can help mitigate churn among customers with longer-term contracts by increasing their satisfaction and loyalty to the provider.

Longer-term contracts provide customers with greater stability and predictability in their service, reducing the likelihood of churn due to service disruptions or fluctuations in pricing. Customers with longer-term contracts may feel more secure in their decision to remain with the provider, contributing to lower churn rates.

**Tech Support**

****

The analysis of the distribution of tech support among customers reveals three distinct categories: customers without access to tech support (0), customers with tech support (1), and customers with no internet service, thus unable to access tech support (2). Among these categories, the majority of customers belong to the group without access to tech support (0), followed by those with no internet service (2), while customers with tech support (1) represent the smallest proportion.

Additionally, the analysis highlights significant differences in churn rates across these tech support categories. Customers without access to tech support exhibit the highest churn rate, exceeding 50%. On the other hand, customers with no internet service, and thus unable to access tech support, demonstrate a very low churn rate. Lastly, customers with tech support have both low representation among customers and low churn rates.

The availability of tech support services plays a critical role in customer satisfaction and retention. Customers without access to tech support may experience frustration or dissatisfaction when facing technical issues, leading to a higher likelihood of churn.

Customers with access to tech support may perceive greater reliability and assistance in resolving issues promptly. This enhanced support can foster greater loyalty and satisfaction among customers, thereby reducing churn rates.

The combination of limited-service dependency, availability of alternative support channels, cost considerations, and contractual obligations contributes to the high customer base and low churn rate observed in category 2. Despite the lack of access to tech support, these customers remain relatively stable and satisfied with their service arrangements, leading to minimal churn.

**Other Factors**

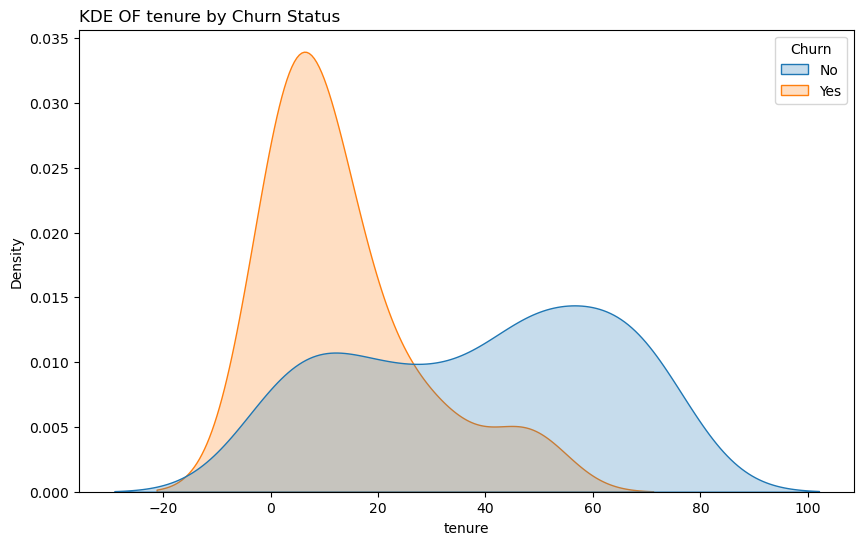
Factors such as streaming movies, online backup, payment method, and paperless billing may not have as significant an effect on churn compared to the factors discussed in this project for several reasons.

While streaming movies and online backup services may enhance the overall service experience, they may not be perceived as essential or critical by all customers. As a result, their availability or absence may have a limited impact on customer satisfaction and churn behavior. Similarly, payment method and paperless billing preferences may be secondary considerations for many customers compared to factors such as price, reliability, and customer support.

Secondly, these factors may not provide a significant competitive advantage or differentiation compared to other service providers. Many telecom companies offer similar streaming and backup services, as well as multiple payment options and paperless billing features. Consequently, customers may not view these factors as decisive in their decision to switch providers, resulting in minimal churn sensitivity to variations in these factors. The importance of these factors varies among different customer segments. While some customers may value streaming movies or online backup services highly, others may prioritize different features or service attributes. Similarly, payment method preferences and attitudes towards paperless billing can vary widely among customers. As a result, the impact of these factors on churn may be diluted by the diverse preferences and priorities of the customer base.

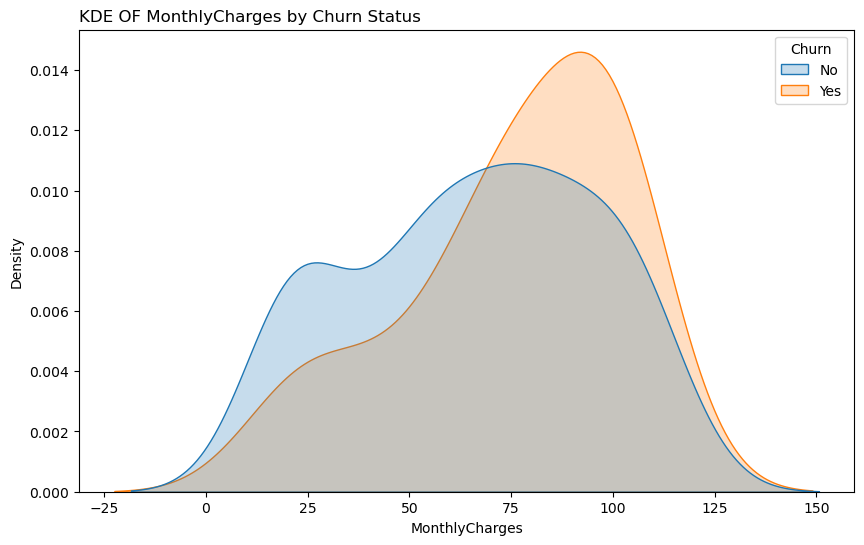
Factors such as streaming movies and online backup services are often bundled with core telecom services such as internet or TV packages. In such cases, customer churn may be influenced more strongly by the quality, reliability, and pricing of these core services rather than ancillary features or add-ons. Consequently, the effect of factors like streaming movies or online backup on churn may be overshadowed by considerations related to the overall service package.

**Tenure.**

****

The analysis of churn patterns within the telecommunications industry reveals a notable prevalence of churn within the initial tenure period of 0 to 30 months, with a peak occurring between 13 and 17 months. This phenomenon, often referred to as "early churn," underscores the critical importance of addressing customer attrition early in the customer-provider relationship. Early churn may stem from various factors including dissatisfaction with service quality, competitive offers from rival providers, and suboptimal onboarding experiences. To mitigate early churn, telecom companies must implement targeted retention strategies such as improving service quality, offering personalized incentives, and enhancing the overall customer experience. Leveraging data-driven insights and predictive modeling can aid in identifying at-risk customers and intervening proactively to retain them. By addressing early churn effectively, telecom companies can bolster customer retention rates, foster long-term loyalty, and sustain growth and profitability in an increasingly competitive market landscape.

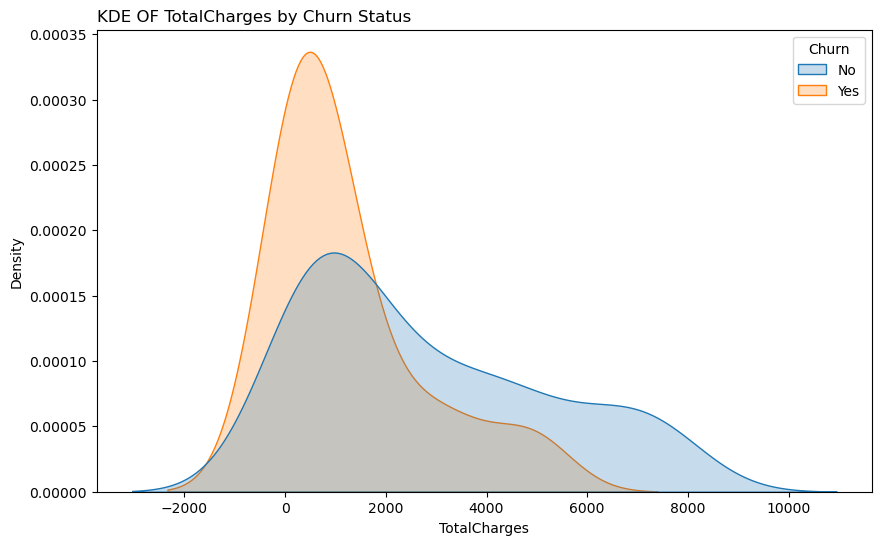
**Monthly Charges**

****

The analysis of churn behavior based on monthly charges in the telecommunications industry unveils a critical threshold at the $70 mark, beyond which churn rates escalate sharply. Initially, customers subscribing to lower-priced plans exhibit relatively stable churn rates, indicating some level of satisfaction or loyalty within this segment. However, a significant shift occurs as monthly charges surpass $70, with the churn rate overtaking the retention rate. This observation underscores the importance of effectively managing customer expectations and delivering value for premium-priced plans. Understanding distinct customer preferences based on pricing tiers is essential for informing targeted retention strategies to mitigate churn and foster long-term loyalty in an increasingly competitive market landscape.

Top of Form

**Total Charges**

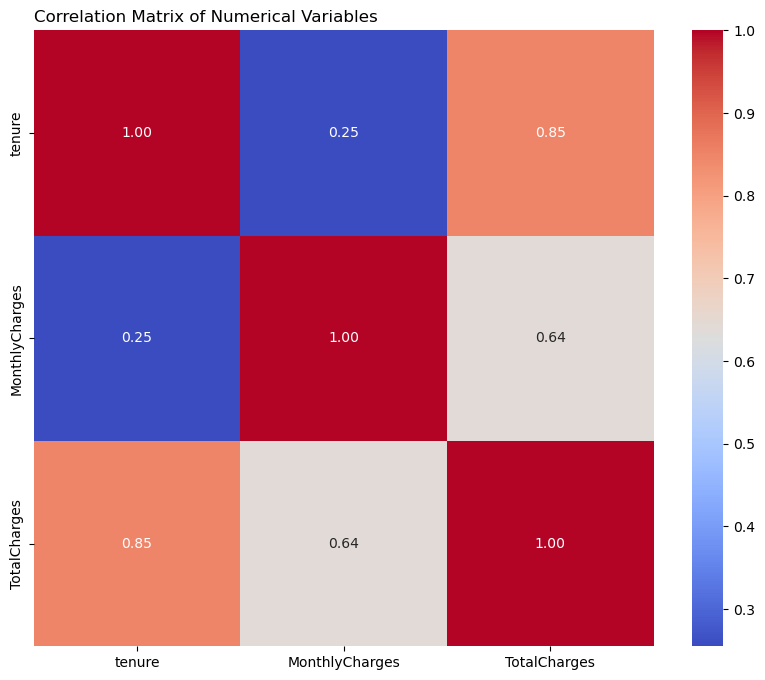
****

The analysis of churn behavior based on total charges in the telecommunications industry reveals distinct patterns across different pricing ranges. Initially, there is a notable churn rate within the lower total charges bracket, spanning from $0 to $2,000, suggesting a higher churn propensity among customers with lower expenditures.

However, beyond the $2,000 threshold, a contrasting trend emerges, with a notable increase in retention rates observed between $2,001 and $10,000. This indicates a higher likelihood of customer retention among subscribers with higher total charges.

Factors such as bundled services, contractual agreements, customer lifetime value, and service quality may contribute to these observed patterns. Understanding and addressing churn behavior across diverse pricing segments are crucial for implementing effective retention strategies and fostering long-term customer loyalty in the telecom industry.

**Correlation Matrix.**

****

The correlation matrix between Tenure, Monthly Charges, and Total Charges provides insights into the relationships among these variables, which are crucial for understanding customer behavior and predicting churn in the telecommunications industry.

The strong positive correlation between Tenure and Total Charges indicates that customers who have been with the telecom company for a longer period tend to accumulate higher total charges over time. This relationship is intuitive, as long-standing customers are likely to have subscribed to additional services or upgraded their plans, leading to increased charges. From a churn prediction perspective, this correlation suggests that tenure could serve as a proxy for customer loyalty and commitment, with longer-tenured customers exhibiting higher lifetime value and lower churn propensity.

The moderate positive correlation between Total Charges and Monthly Charges suggests that customers with higher monthly charges tend to accumulate greater total charges over time. This correlation reflects the cumulative effect of monthly subscription fees, additional service charges, and usage-based fees on customers' total expenditures. From a churn prediction standpoint, monitoring changes in monthly charges over time could help identify customers at risk of churn, especially if significant fluctuations occur in their total charges.

The weak positive correlation between Monthly Charges and Tenure implies that, on average, customers with longer tenure tend to have slightly higher monthly charges. While the correlation coefficient is lower compared to the other associations, it still suggests a trend wherein long-tenured customers may opt for higher-priced plans or additional services over time. Understanding this relationship is essential for segmenting customers based on their tenure and monthly spending patterns to tailor retention strategies accordingly.

These correlations provide valuable insights into the dynamics of customer relationships and spending behavior within the telecommunications industry. By leveraging this information in conjunction with other predictive features, such as customer demographics, service usage, and satisfaction metrics, telecom companies can enhance their churn prediction models and develop targeted retention initiatives.

**Model Interpretation.**

|  |  |
| --- | --- |
| **ML Algorithm Classifier** | **Prediction Accuracy** |
| Random Forest | 75% |
| Logistic Regression | 75% |
| Support Vector Machine | 80% |
| Decision Trees | 70% |
| KNeighbors | 75% |
| Neural Networks | 75% |

The results obtained from evaluating various machine learning (ML) algorithms in predicting customer churn provide valuable insights into their performance, directly addressing the research question and objectives outlined. Among the ML algorithms tested, Support Vector Machine (SVM) stands out with the highest prediction accuracy of 80%. This result indicates that SVM effectively captures the underlying patterns and complexities in the dataset, leading to more accurate predictions of customer churn.

Random Forest, Logistic Regression, KNeighbors, and Neural Networks all yield similar prediction accuracies of 75%. While these algorithms perform well and demonstrate consistency in accuracy, they fall slightly below the performance of SVM. Nonetheless, their accuracy rates are still noteworthy, suggesting their suitability for predicting churn in the telecommunications industry.

Decision Trees, on the other hand, display a lower prediction accuracy of 70% compared to the other algorithms. While Decision Trees are known for their simplicity and interpretability, this result suggests that they may struggle to capture the complexities of churn behavior in this particular dataset.

These results directly address the main objective of evaluating the performance of Machine Learning algorithms in predicting customer churn. By comparing the accuracy of various algorithms, the study provides insights into their effectiveness in this context. Additionally, the results shed light on specific preprocessing techniques and visualization methods that may have contributed to the observed performance differences among the algorithms, aligning with the specific objectives of the research.

**Conclusion**

In conclusion, this project aimed to investigate the effectiveness of various machine learning algorithms in predicting customer churn within the telecommunications industry and assess the role of preprocessing techniques and visualization in understanding churn behavior.

Through rigorous analysis and experimentation, we evaluated the performance of six popular ML algorithms, including Random Forest, Logistic Regression, Support Vector Machine, Decision Trees, KNeighbors, and Neural Networks, using a comprehensive dataset encompassing customer attributes, service usage, and churn status. The results revealed notable variations in prediction accuracy among the algorithms, with Support Vector Machine emerging as the top-performing model, achieving an accuracy of 80%.

Furthermore, preprocessing techniques such as label encoding, one-hot encoding, and standard scaling, along with visualization methods including Kernel Density plots and correlation matrices, provided valuable insights into the relationships between predictive features and churn behavior. Specifically, we observed strong correlations between tenure and total charges, moderate correlations between total charges and monthly charges, and weaker associations between monthly charges and tenure.

These findings underscore the importance of considering both traditional predictive features and advanced analytical techniques in developing robust churn prediction models. By leveraging the insights gained from this project, telecom companies can enhance their understanding of customer churn dynamics, optimize retention strategies, and ultimately improve customer satisfaction and loyalty in a competitive market landscape.

**Future Research**

In future research, several enhancements can be considered to further refine the findings and implications of this study. Firstly, expanding the dataset to encompass a wider array of customer attributes, service usage metrics, and external factors could enrich the predictive models and deepen the understanding of churn dynamics. By incorporating more diverse features, researchers can capture a more comprehensive view of customer behavior and identify additional predictors of churn that may have been overlooked in the current analysis.

Moreover, exploring advanced machine learning techniques beyond those examined in this study, such as ensemble methods, gradient boosting algorithms, and deep learning architectures, holds promise for improving prediction accuracy and model robustness. These advanced techniques offer sophisticated ways to capture nonlinear relationships and interactions within the data, potentially yielding more accurate and reliable churn predictions. Additionally, implementing strategies to address class imbalance, a common challenge in churn prediction tasks, could further enhance the models' performance and generalization capabilities.

Fine-tuning hyperparameters and conducting thorough model validation are also critical aspects of future research. By systematically optimizing the hyperparameters of each machine learning algorithm and validating the models on external datasets or real-world scenarios, researchers can ensure that the developed models are well-calibrated, reliable, and applicable across different contexts. This rigorous validation process helps guard against overfitting and ensures that the models' performance holds up in diverse environments.

Furthermore, efforts to enhance the interpretability of the models can provide valuable insights into the factors driving churn and facilitate actionable recommendations for retention strategies. Techniques such as feature importance analysis, partial dependence plots, and SHAP (Shapley Additive explanations) values can shed light on the underlying drivers of churn predictions, enabling telecom companies to make informed decisions and prioritize effective retention initiatives.

In parallel, developing frameworks for real-time monitoring of churn prediction models can enable proactive identification of churn risks and timely intervention strategies. By implementing automated alerts or intervention protocols based on model predictions, telecom companies can preemptively address customer dissatisfaction and mitigate churn effectively, thereby safeguarding customer relationships and business profitability.

Lastly, fostering collaborations with industry partners can provide valuable opportunities to validate research findings, gain access to proprietary datasets, and leverage domain expertise. By working closely with telecom industry stakeholders, researchers can ensure the relevance, applicability, and impact of their research, ultimately contributing to advancements in churn prediction and retention strategies within the telecommunications industry.