Kaylyn Dressel

ISOM 835 2 Predictive Analytics

December 12, 2025

Final Predictive Analytics Assignment

Executive Summary:

Customer churn presents a significant challenge for subscription-based businesses, where retaining existing customers is often more cost-effective than acquiring new ones. This project applies predictive analytics to identify customers at high risk of churn and to uncover the key factors influencing customer attrition.

The analysis uses the Telco Customer Churn dataset, which contains approximately 7,000 customer records with information on tenure, service subscriptions, billing details, and contract types. Exploratory analysis revealed that churn is strongly associated with shorter tenure, month-to-month contracts, and higher monthly charges.

Two predictive models were developed and evaluated: a baseline Logistic Regression model and a more advanced Random Forest model. The tuned Random Forest model outperformed the baseline across key evaluation metrics, particularly in identifying customers likely to churn.

Based on these findings, the analysis recommends focusing retention efforts on early-stage customers, incentivizing longer-term contracts, and addressing pricing sensitivity among high-cost customers. Overall, this project demonstrates how predictive analytics can support proactive, data-driven customer retention strategies.

Introduction & Business Context:

Customer churn is a critical business issue for subscription-based organizations, particularly in competitive industries such as telecommunications. When customers leave, companies lose recurring revenue and incur additional costs related to customer acquisition and onboarding. As a result, understanding churn behavior and identifying customers at risk is essential for sustaining long-term profitability.

Retention is especially important because customers who remain longer generate greater lifetime value and contribute to more stable revenue streams. High churn rates can also signal broader issues related to pricing, service quality, or customer engagement. Rather than responding after customers leave, organizations increasingly rely on data-driven approaches to anticipate churn and intervene proactively.

The objective of this project is to use predictive analytics to identify customers most likely to churn and to understand the factors driving those outcomes. By analyzing historical customer data, the project aims to support more effective retention strategies and improved decision-making.

This analysis addresses the following research questions:

1. Which customer characteristics are most strongly associated with churn?
2. Can churn be predicted with reasonable accuracy using historical data?
3. How do contract structures and pricing levels influence churn risk?
4. Which customer segments should be prioritized for retention efforts?

The dataset used in this project is the Telco Customer Churn dataset sourced from Kaggle. It contains approximately 7,000 customer records with a mix of numeric and categorical variables related to customer tenure, services, billing, and contract terms. The target variable indicates whether a customer has churned, making the dataset well-suited for predictive classification analysis.

Exploratory Data Analysis:

Data Structure and Characteristics

The dataset contains approximately 7,000 customer records with a mix of numeric and categorical features describing customer tenure, service usage, billing, and contract details. The target variable, churn, indicates whether a customer discontinued service. As expected, the dataset is moderately imbalanced, with more customers remaining than churning.

Key Patterns and Relationships Discovered

Customer tenure shows a strong relationship with churn. Customers who churned generally had significantly shorter tenure than those who remained, indicating that churn risk is highest early in the customer lifecycle.

Monthly charges also exhibit a meaningful relationship with churn. Customers who churned tended to have higher and more variable monthly charges, suggesting pricing sensitivity or perceived value concerns.

Contract type emerged as one of the strongest churn indicators. Customers on month-to-month contracts showed substantially higher churn rates compared to those on one- or two-year contracts, highlighting the stabilizing effect of longer-term commitments.

Internet service type further differentiated churn behavior. Customers using fiber optic service experienced higher churn rates than those using DSL or no internet service, potentially reflecting higher expectations or competitive alternatives.

Correlation analysis of numeric features showed a strong positive relationship between tenure and total charges, which is expected given cumulative billing over time.

Data Quality Issues Encountered

Several data quality issues were identified during exploratory analysis. The most notable issue involved the TotalCharges variable, which was stored as a text field rather than a numeric variable. Some entries contained blank or invalid values, which required conversion to numeric format and treatment as missing values. This issue was addressed during preprocessing.

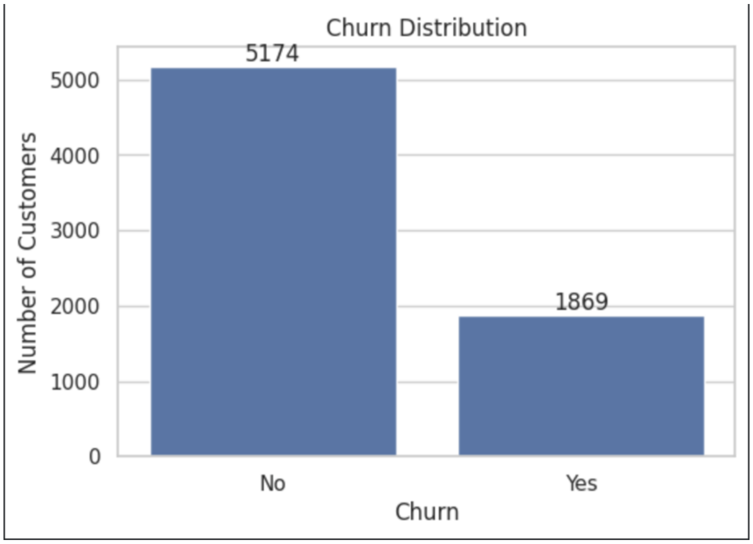
Aside from this, the dataset was largely complete, with minimal missing data across most features. No significant duplicate records were identified. Some categorical variables contained multiple similar levels (e.g., service options labeled “Yes,” “No,” or “No internet service”), which required careful handling during preprocessing to ensure consistent interpretation.

Overall, while minor cleaning was required, the dataset was of relatively high quality and suitable for both exploratory analysis and predictive modeling.

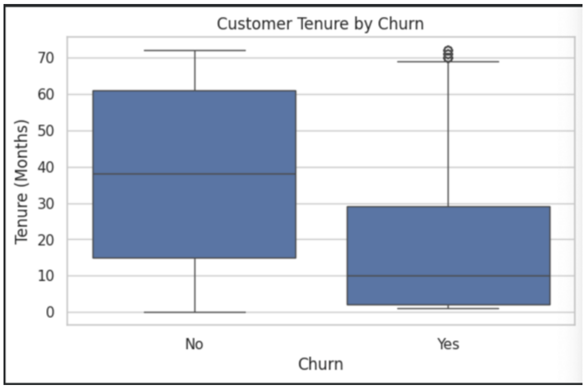
Visualizations and Interpretations

Multiple visualizations were used to support exploratory analysis and to clearly communicate findings. Charts and interpretations are listed below.

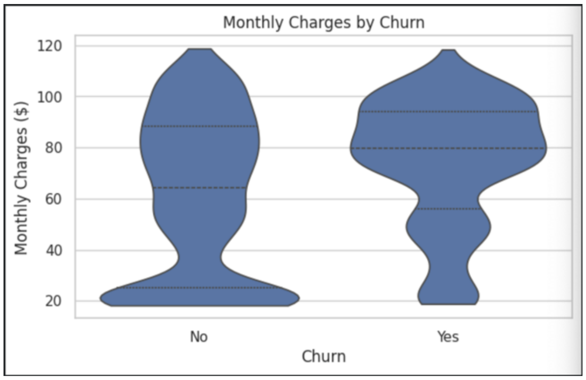
This bar chart of the churn distribution illustrated that most customers do not churn, but a meaningful minority do. This confirmed that churn prediction is a relevant and non-trivial problem for the business.



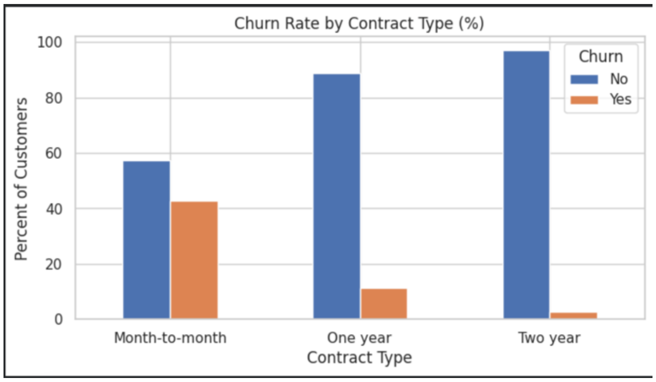
The box plot comparing tenure by churn status showed that customers who churn typically have much shorter tenure. This visualization reinforced the importance of early-stage customer engagement and suggested that retention strategies should focus on new customers.



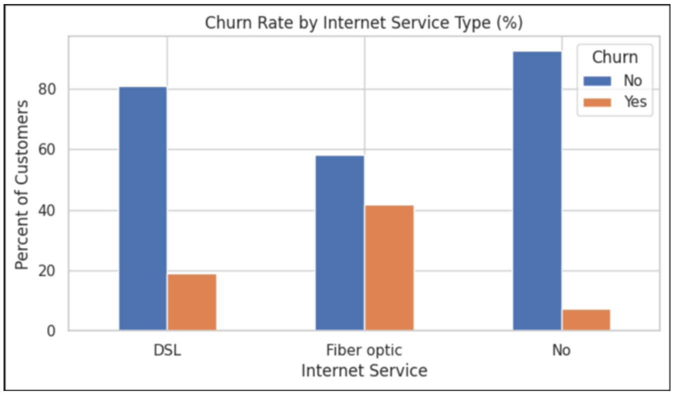
This violin plot of monthly charges by churn status revealed that churned customers tend to have higher monthly charges and a wider range of billing amounts. This suggests that pricing and cost perception may influence churn decisions.



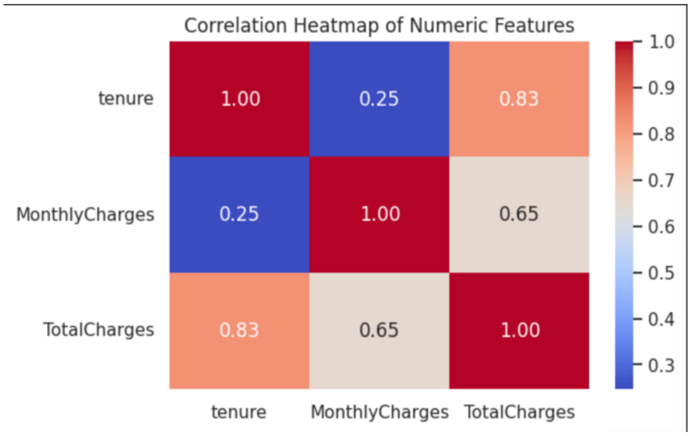
The stacked bar chart comparing churn rates across contract types highlighted a stark contrast between month-to-month customers and those on longer-term contracts. Month-to-month customers had the highest churn rates, while two-year contracts showed the lowest churn, indicating that contract commitment plays a major role in retention.



This is another stacked bar chart examining internet service type showed that fiber optic customers experienced higher churn rates than other groups. This finding suggests the need for closer examination of service quality, pricing, or competition within this segment.



Finally, this correlation heatmap of numeric features illustrated expected relationships, particularly the strong correlation between tenure and total charges. While correlation does not imply causation, this visualization helped confirm the internal consistency of the data and guided feature selection for modeling.



EDA Summary

The exploratory data analysis provided clear evidence that churn is closely linked to customer tenure, contract structure, and billing levels. Customers who are early in their relationship with the company, those on flexible contract terms, and those facing higher monthly charges are at the greatest risk of churn. These findings informed subsequent preprocessing decisions and guided the selection of predictive modeling techniques.

By identifying these patterns early, the analysis established a strong foundation for building predictive models and developing actionable business recommendations focused on customer retention.

Methodology:

Data Preprocessing Steps

Before building predictive models, the dataset was prepared to ensure consistency, accuracy, and reliability. Initial preprocessing focused on correcting data types and handling missing values. One variable, TotalCharges, was stored as a text field and contained blank entries. This variable was converted to a numeric format, with invalid values treated as missing. A customer identifier field was removed, as it serves only as a unique label and does not provide predictive value.

The dataset includes both numeric and categorical variables, requiring different preprocessing approaches. Numeric features were prepared by imputing missing values using the median and standardizing values to ensure comparable scales across variables. Categorical features were prepared by imputing missing values using the most common category and converting them into numerical form using one-hot encoding. This approach allows models to recognize differences between customer groups without assuming any inherent ordering among categories.

All preprocessing steps were implemented using a unified pipeline approach. This ensured that the same transformations were applied consistently across training and testing data and reduced the risk of data leakage. The dataset was then split into training and testing subsets using stratified sampling to preserve the original proportion of churned and non-churned customers.

Feature Engineering Decisions

Feature engineering decisions were guided by exploratory analysis and business relevance. Most original features were retained, as they captured key aspects of customer behavior, including tenure, contract structure, billing information, and service subscriptions.

Customer tenure received particular attention, as exploratory analysis indicated strong differences in churn risk across different stages of the customer lifecycle. While the primary models used tenure as a continuous variable, tenure-based groupings were explored to support business interpretation of churn patterns among early-stage versus long-term customers.

Categorical variables such as contract type, internet service, and payment method were preserved and encoded to ensure that the models could capture meaningful differences across customer segments. Feature engineering was intentionally kept simple to balance predictive performance with interpretability, ensuring that model results could be clearly communicated to non-technical stakeholders.

Models Selected and Rationale

Two predictive models were selected to evaluate customer churn: Logistic Regression and Random Forest. These models were chosen to provide both a baseline benchmark and a more flexible, advanced modeling approach.

Logistic Regression was used as the baseline model due to its simplicity and interpretability. It is commonly applied in churn prediction and provides a transparent reference point for understanding how customer attributes relate to churn risk. While its assumptions limit its ability to capture complex relationships, it serves as a useful benchmark for evaluating more advanced techniques.

Random Forest was selected as the advanced model because it can capture non-linear relationships and interactions between customer attributes. This makes it well-suited for modeling complex behavioral patterns that are common in customer data. In addition, Random Forest produces feature importance measures, which support business interpretation and help identify the most influential drivers of churn.

Using both models allowed for a clear comparison between interpretability and predictive performance.

Evaluation Metrics Chosen

Model performance was evaluated using several classification metrics, with primary emphasis on ROC-AUC. ROC-AUC measures a model's ability to distinguish between churned and non-churned customers across a range of decision thresholds. This metric is particularly appropriate for churn prediction because it is not dependent on a single cutoff value and performs well in the presence of class imbalance.

Additional metrics, including precision, recall, and confusion matrices, were also considered. Recall was especially important, as failing to identify a customer who is likely to churn represents a missed opportunity for retention intervention. Precision provided insight into how efficiently the model identifies true churners without over-targeting customers who are unlikely to leave.

Together, these metrics provided a balanced assessment of both predictive accuracy and business usefulness.

Hyperparameter Tuning Approach

To improve the performance of the Random Forest model, hyperparameter tuning was conducted using a randomized search approach. Rather than testing every possible parameter combination, randomized search evaluates a subset of configurations, allowing for efficient optimization while managing computational cost.

Key parameters explored included the number of trees, maximum tree depth, and minimum sample thresholds for splits and leaf nodes. Cross-validation was used during tuning to ensure that performance improvements generalized beyond a single data split.

The tuned Random Forest model demonstrated improved performance compared to both the baseline Logistic Regression and the untuned Random Forest model, confirming that optimization meaningfully enhanced the model’s ability to identify customers at risk of churn.

Results and Model Comparison:

Model Performance Overview

Two predictive models were evaluated to assess their ability to identify customers at risk of churn: a baseline Logistic Regression model and an advanced Random Forest model. The Random Forest model was also evaluated after hyperparameter tuning. Model performance was assessed using ROC-AUC as the primary metric, along with precision, recall, confusion matrices, and ROC curves. These metrics were selected to balance overall predictive accuracy with business relevance, particularly the ability to identify customers likely to churn.

Overall, all models performed better than random guessing, confirming that churn can be reasonably predicted using historical customer data. However, meaningful differences emerged in each model’s ability to correctly identify churned customers.

Logistic Regression Results

Logistic Regression was used as a baseline due to its interpretability and widespread use in churn prediction. The model achieved a moderate ROC-AUC score, indicating some ability to distinguish between churned and non-churned customers. The confusion matrix shows that the model performed well at identifying customers who did not churn but was less effective at identifying customers who ultimately left.

This imbalance is reflected in lower recall for churned customers. From a business perspective, this limitation is significant because missed churners represent lost opportunities for targeted retention efforts. While Logistic Regression provides useful directional insights and transparency, its performance suggests that more flexible models may better capture complex customer behavior.

Random Forest Results

The Random Forest model demonstrated improved performance compared to Logistic Regression across key metrics. The untuned Random Forest achieved a higher ROC-AUC score, indicating better overall separation between churned and non-churned customers. Its confusion matrix shows improved recall for churned customers, meaning the model was more effective at identifying customers at risk of leaving.

Although this improvement came with a modest trade-off in precision, the increase in correctly identified churners is valuable in a retention context, where outreach strategies can be refined further. The ROC curve for the Random Forest model consistently lies above that of the Logistic Regression model, confirming stronger performance across a range of classification thresholds.

Hyperparameter Tuning Results

To further enhance model performance, the Random Forest model was optimized using hyperparameter tuning. The tuned Random Forest achieved the highest ROC-AUC score among all models evaluated. Improvements were observed in both recall and overall discrimination, indicating that tuning helped the model better capture underlying churn patterns.

Cross-validation during tuning helped ensure that performance gains generalized beyond a single train-test split. These results demonstrate that model optimization meaningfully improved predictive capability and reinforced the value of using an advanced, tuned approach for churn prediction.

Model Comparison

When comparing all three models (Logistic Regression, Random Forest, and tuned Random Forest) the tuned Random Forest consistently delivered the strongest overall performance. Logistic Regression provided interpretability but lagged in identifying churned customers. The untuned Random Forest improved recall and discrimination, while the tuned Random Forest offered the best balance between accuracy and business usefulness.

From a business standpoint, the tuned Random Forest's improved recall is particularly important. Identifying a higher proportion of customers at risk of churn allows the organization to allocate retention resources more effectively, even if some customers flagged by the model do not ultimately churn.

Feature Importance Analysis

Feature importance analysis from the Random Forest model provided insight into the key drivers of churn. Customer tenure emerged as the most influential feature, reinforcing earlier findings that churn risk is highest early in the customer lifecycle. Contract type was another major driver, with month-to-month contracts strongly associated with higher churn risk. Monthly charges also ranked highly, suggesting that pricing sensitivity or perceived value plays a role in churn decisions.

Additional service-related variables, such as internet service type and payment method, contributed to churn predictions but to a lesser extent. These results align closely with patterns identified during exploratory data analysis and support actionable business interpretation.

Best Model Selection and Justification

Based on the comparative analysis, the tuned Random Forest model was selected as the best-performing model. This decision was driven by its superior ROC-AUC score, improved recall for churned customers, and ability to provide interpretable feature importance insights. While

Logistic Regression offered greater transparency, its lower effectiveness in identifying churned customers limited its practical value.

The tuned Random Forest strikes the strongest balance between predictive performance and business insight. In a real-world setting, this model would provide the greatest value by enabling proactive identification of high-risk customers while supporting informed retention strategies. These results form the foundation for the business recommendations presented in the following section.

Business Insights & Recommendations:

Translating Analytical Results into Business Value

The predictive analysis demonstrates that customer churn is driven by clear and actionable factors rather than random behavior. By analyzing customer tenure, contract type, and billing information, the models can identify customers at high risk of churn early enough for meaningful intervention. This enables the business to move from reactive churn management to proactive retention strategies, improving customer lifetime value and reducing revenue loss.

The strongest churn drivers (short tenure, month-to-month contracts, and higher monthly charges) align directly with business levers that can be addressed through targeted engagement, pricing strategies, and contract incentives. As a result, the analytical findings translate directly into practical opportunities to improve retention outcomes.

Actionable Recommendations

Focus on Early-Customer Retention: Customers in the early stages of their relationship with the company are most likely to churn. The business should prioritize onboarding and early engagement efforts, such as proactive outreach, service education, or introductory offers, to reduce early dissatisfaction and improve long-term retention.

Encourage Longer-Term Contracts: Month-to-month contracts are associated with significantly higher churn. Offering incentives such as discounted pricing, bundled services, or loyalty benefits can encourage customers to transition to longer-term agreements, increasing stability and reducing churn risk.

Address Pricing Sensitivity: Customers with higher monthly charges show elevated churn risk. Targeted pricing reviews or value-based offers for these customers may improve perceived value and reduce the likelihood of churn.

Use Predictive Insights to Target Outreach: Churn risk scores can be used to prioritize retention efforts. Rather than applying uniform outreach strategies, teams can focus resources on high-risk customers, improving efficiency and the effectiveness of retention campaigns.

Implementation Considerations and Expected Impact

To implement these recommendations, churn predictions should be integrated into existing customer management systems and refreshed regularly to reflect changes in customer behavior. Clear ownership across marketing, customer support, and account management teams will be critical to ensure insights translate into action. Retention efforts should be positioned as customer-centric and value-driven to maintain positive customer experiences.

By focusing retention efforts on the highest-risk segments, the organization can reduce churn, increase customer lifetime value, and improve the return on retention investments. Over time, this data-driven approach supports stronger customer relationships and more informed strategic decision-making.

Ethics & Responsible AI:

Potential Biases Identified

The churn prediction models can reflect biases that already exist in the data. Factors like contract type, payment method, or pricing level may indirectly relate to customers' financial situations, which could cause certain groups to be labeled as higher risk more often than others. Additionally, because the model is based on past behavior, it may not fully account for future shifts in customer preferences or changes in the market.

Fairness Considerations

To promote fairness, it's important to regularly review model results across different customer segments to make sure no group is being unfairly impacted. Churn predictions should be used to guide supportive actions like personalized outreach or retention offers, rather than to penalize or limit customers. Keeping the focus on positive, value-driven interventions helps reduce the risk of unintended bias or unfair treatment.

Privacy and Security Implications

The dataset includes sensitive customer billing and service information, making data protection essential. Access to both the data and model outputs should be limited to authorized personnel, and appropriate security measures should be in place to protect customer privacy. Only data necessary for prediction should be used to minimize privacy risk.

Recommendations for Responsible Deployment

Churn prediction models should support customer experience and retention, not automated decision-making without oversight. Regular monitoring and retraining are recommended to maintain accuracy and reduce bias over time. Human judgment should remain central to how model outputs are used, ensuring responsible and ethical application.

Conclusion & Future Work:

Summary of Achievements

This project demonstrated how predictive analytics can be used to better understand and predict customer churn in a subscription-based business. Through exploratory analysis, data preprocessing, and modeling, key drivers of churn were identified, including customer tenure, contract type, and monthly charges. Two models were evaluated, with the tuned Random Forest model providing the strongest overall performance. Most importantly, the analysis translated model outputs into actionable insights that can support proactive customer retention strategies.

Limitations of the Current Approach

Several limitations should be acknowledged. The analysis relied on a single historical dataset, which may not fully reflect future customer behavior or changing market conditions. The dataset also lacked qualitative factors such as customer satisfaction or service quality, which could influence churn decisions. Additionally, while the selected model improved churn identification, false positives and false negatives remain unavoidable in any predictive approach.

Suggestions for Future Improvements

Future work could enhance the analysis by incorporating additional data sources, such as customer support interactions, usage patterns, or survey feedback. More advanced modeling techniques or time-based approaches could also be explored to improve predictive accuracy. In a real-world deployment, continuous monitoring and periodic retraining would be necessary to maintain model relevance.

Lessons Learned

This project reinforced the importance of aligning predictive analytics with business objectives. Strong preprocessing, appropriate evaluation metrics, and clear interpretation are critical to building useful models. Most importantly, predictive analytics delivers the greatest value when used to inform thoughtful, customer-focused decision-making.

References & Acknowledgments:

Kaggle. (n.d.). Kaggle datasets. https://www.kaggle.com/datasets

scikit-learn developers. (n.d.). scikit-learn: Machine learning in Python. https://scikit-learn.org/stable/

Codefinity. (n.d.). Google Colab tutorial. https://codefinity.com/blog/Google-Colab-Tutorial

Git documentation. (n.d.). Git tutorial. https://git-scm.com/docs/gittutorial

TutorialsPoint. (n.d.). Google Colab tutorial. https://www.tutorialspoint.com/google_colab/index.htm

GitHub. (n.d.). Introduction to GitHub. https://github.com/skills/introduction-to-github

HubSpot. (n.d.). Git and GitHub tutorial for beginners. https://product.hubspot.com/blog/git-and-github-tutorial-for-beginners

Altair developers. (n.d.). Altair: Declarative visualization in Python. https://altair-viz.github.io/getting_started/overview.html

OpenAI. (2025). ChatGPT (GPT-based language model). https://www.openai.com/