Q1. Clearly state the business case (who/problem/solution/payoff matrix).

This analysis centers on a pivotal challenge for a telecommunications company: accurately forecasting customer churn. Churn prediction is vital as retaining customers is typically more cost-effective than acquiring new ones, and high churn rates can negatively impact both profitability and market reputation. To tackle this issue, the company plans to employ a variety of advanced analytics techniques, including logistic regression, decision trees, random forests, and gradient boosting models. These methods aim to predict which customers are most likely to churn by analyzing extensive customer data. The project's objective is to pinpoint key churn predictors, thereby enabling the company to customize its retention strategies more effectively. The ultimate goals are to improve customer engagement and loyalty programs, reduce churn rates, and optimize marketing and operational resources, which should contribute positively to the company's financial health. The effectiveness of these efforts will be assessed using model performance metrics such as Recall, Precision, F1 Score, ROC AUC, and particularly through LogLoss, which directly relates to the predictive certainty of the models. Additionally, a financial metric, the Maximum Payoff, will estimate the economic impact of various retention campaigns. This dual focus on model accuracy and financial outcomes promises to deliver strategic insights, guiding the allocation of resources toward the most effective retention measures and thus securing a higher ROI and a more competitive stance in the telecommunications sector.

The first step in this process will be an exploratory data analysis of the "telco_churn" dataset to understand the data's characteristics and patterns. This will be followed by data preprocessing to prepare it for modeling. Subsequently, several predictive models will be developed and evaluated for their ability to forecast customer churn, with a comparative analysis of their performance based on the prioritized metrics, including LogLoss. The final phase will involve a detailed examination of the primary churn predictors identified by the most effective model, with a focus on visualizing their impacts and offering actionable recommendations for each.

Financial Assumptions:

- ***Marketing Effort Costs (MEC):*** This cost refers to the expenditure on marketing campaigns aimed at retaining customers or preventing churn. It includes the cost of communications, promotions, and other activities designed to engage customers who are identified as at risk of leaving. For this analysis, we assume an MEC of -$150 per customer, reflecting the direct cost that will be subtracted from any revenue generated from retaining customers (TPs) or wasted on customers not at risk (FPs).
- ***Service Discounts (SD):*** To incentivize customers to stay, companies often offer discounts or additional services. These can effectively reduce churn but also represent a cost to the company. We've assumed an SD of -$200 for each retention effort that results in a TP. This discount reduces the immediate revenue from the retained customer, but it is an investment in future revenue stability.
- ***Customer Service Costs (CSC):*** This cost covers the resources used during customer service interactions, such as time spent by customer service representatives, the use of support systems, and other overheads. For this example, we assume a CSC of -$50. It's applied to each TP when customer service is involved in the retention effort, and it could

also be relevant to FNs if there's an attempt to re-engage customers who have already churned.

- ***Opportunity Costs (OC):*** The Opportunity Cost (OC) in the analysis relates to missed revenue opportunities due to incorrect model predictions. Specifically, for False Negatives (FN), where the model fails to identify customers at risk of churning who then proceed to churn, OC encapsulates
- ***Churn Recovery Costs (CRC):*** These are costs associated with efforts to win back customers who have already churned. Recovery tactics might include special offers, dedicated customer service, and other re-engagement strategies. In our scenario, we assume a CRC of -$250 for each FN, reflecting the additional costs the company incurs in attempts to regain the customer's business.

A breakdown the payoff matrix in this scenario would be the following:

- ***True Positives (TP):*** Customers correctly identified as at risk of churning, where retention efforts are successful. The payoff for TP would be the saved revenue from preventing a churn.
    - Hypothetical Value: $600 ($1000 retention value - $150 MEC - $200 SD - $50 CSC)
- ***False Positives (FP):*** These are instances where the model incorrectly predicts churn or subscription. The cost here involves wasted resources on unnecessary retention efforts or ineffective marketing campaigns targeting the wrong individuals.
    - Hypothetical Value: -$150 (MEC)
- ***True Negatives (TN):*** Customers correctly identified as not at risk of churning, requiring no action. There's no direct cost or revenue associated with TNs, but correctly identifying TNs helps avoid unnecessary expenditures on retention efforts.
    - Hypothetical Value: $0
- ***False Negatives (FN):*** These are situations where the model fails to identify a potential churn. The cost is the missed opportunity for revenue, either by losing a customer or not retaining the subscriber with recovery tactics.
    - Hypothetical Value: -$1550 (-$1000 lost revenue - $300 OC - $250 CRC)

Q2. Perform exploratory data analysis, pre-process the data as necessary.

The dataset contains various features related to customer account information and service usage, along with the target variable churn which indicates whether a customer has left the company.

- ***Categorical Variables:*** We need to ensure that state, area_code, international_plan, voice_mail_plan, and churn are coded as categorical.
- ***Numerical Variables:*** The rest, including account_length, number_vmail_messages, various call minutes and charges, total_intl_minutes, total_intl_calls, total_intl_charge, and number_customer_service_calls would be coded as numerical.
- ***Missing Values:*** There are no missing values in any of the columns, which simplifies the preprocessing step as we don't need to impute missing data.
- ***Variance of Features:*** All features exhibit some level of variance, with no indications of features having zero variance. This means each feature potentially contributes

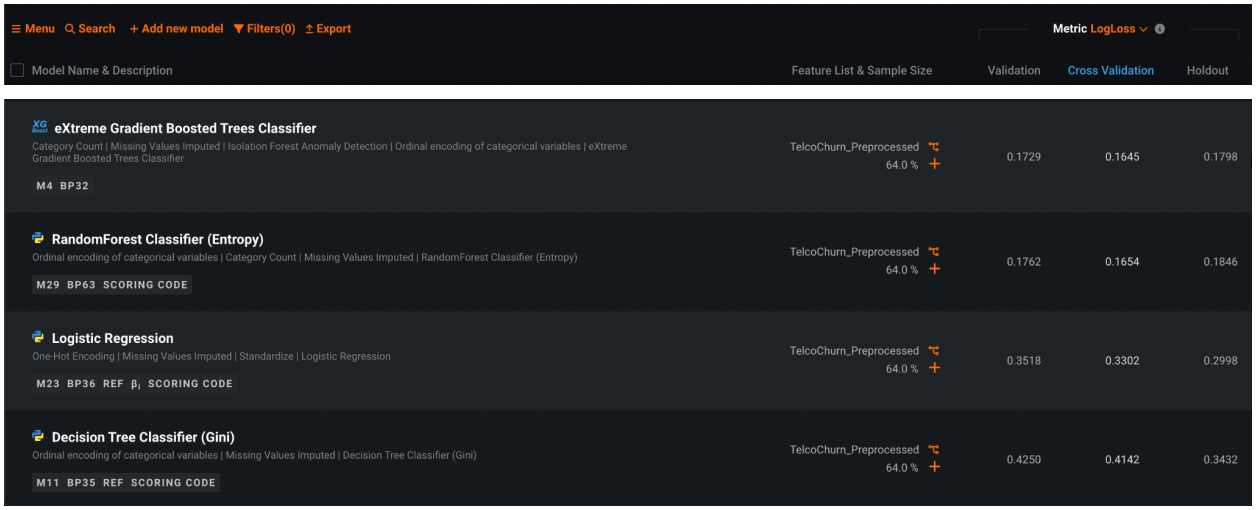
information useful for predicting customer churn. Features like international_plan, voice_mail_plan, and the various one-hot encoded state and area_code variables have lower variance compared to continuous numerical features like total_day_minutes, suggesting varying degrees of informativeness.

- ***High Cardinality Features:*** The "state" feature, including Washington D.C., exhibits a cardinality of 51, leading to its transformation into 51 distinct binary columns. Such high cardinality introduces a layer of complexity to the predictive model and raises the risk of overfitting. Since this was a low impact feature, I omitted it during the pre-processing step to avoid overfitting.
- ***Outliers:*** The features total_day_calls, total_eve_calls, and total_night_calls have outliers, suggesting that a minority of customers engage in call activities (whether during the day, evening, or night) to an extent significantly different from the general population. This variance indicates diverse customer behavior patterns, which, while potentially skewing aggregate metrics, could offer valuable insights into specific usage trends and customer needs. In this analysis, I will not be modifying these outliers, as the analysis benefits from a more nuanced and comprehensive dataset. This enables a more accurate and meaningful interpretation of customer behavior.
- ***Checking for 100% Redundancies:*** total_day_minutes and total_day_charge are likely to be highly correlated since charges are usually a function of the number of minutes. Similarly, total_eve_minutes and total_eve_charge, total_night_minutes and total_night_charge, and total_intl_minutes and total_intl_charge would also be expected to have high correlations for the same reason. For this analysis, we have to identify 100% redundant features, so we need to check for columns that are either duplicates of each other or have a constant value across all rows. Exploratory data analysis reveals that there are no 100% redundant features in the dataset. This suggests that there are no features to exclude immediately based on being 100% redundant.

Feature Name	Data Quality	Index	Importance	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
total_day_minutes		7		Numeric	1,682	0	180	54.24	180	0	347
total_day_charge		9		Numeric	1,682	0	30.57	9.22	30.67	0	58.96
number_customer_service_calls		19		Numeric	10	0	1.57	1.31	1	0	9
international_plan		4		Categorical	2	0					
voice_mail_plan		5		Categorical	2	0					
number_vmail_messages		6		Numeric	46	0	7.73	13.53	0	0	52
total_intl_calls		17		Numeric	21	0	4.42	2.46	4	0	20
total_intl_charge		18		Numeric	161	0	2.78	0.74	2.81	0	5.40
total_intl_minutes		16		Numeric	161	0	10.28	2.75	10.40	0	20
total_eve_minutes		10		Numeric	1,613	0	200	50.16	201	0	359
total_eve_charge		12		Numeric	1,439	0	17.01	4.26	17.06	0	30.54
total_night_minutes		13		Numeric	1,597	0	201	50.12	201	23.20	382
total_night_charge		15		Numeric	941	0	9.03	2.26	9.05	1.04	17.19
state		1		Categorical	51	0					
area_code		3		Categorical	3	0					
total_day_calls		8		Numeric	118	0	99.68	19.95	100	0	165

total_eve_calls		11		Numeric	120	0	100	20.02	100	0	170
account_length		2		Numeric	212	0	101	39.70	100	1	243
total_night_calls		14		Numeric	124	0	99.77	19.99	100	33	170

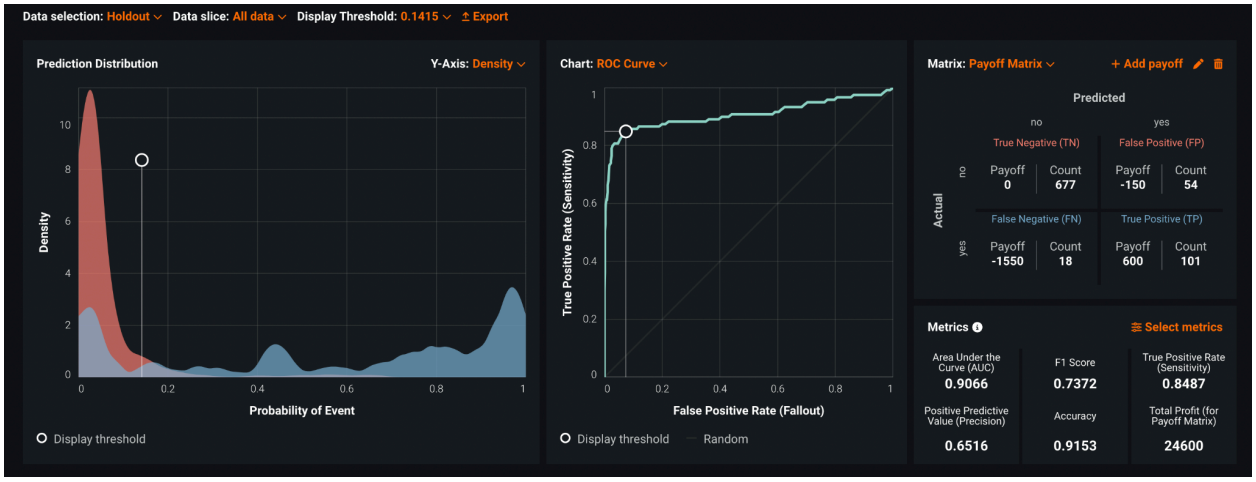
Q3. Make use of all the modeling techniques that you know to build models to forecast customer churn. Show a table with a summary of model performance metrics: Recall, Precision, F1, ROC AUC, max payoff. Identify the best performing model. Explain your choice of metric.



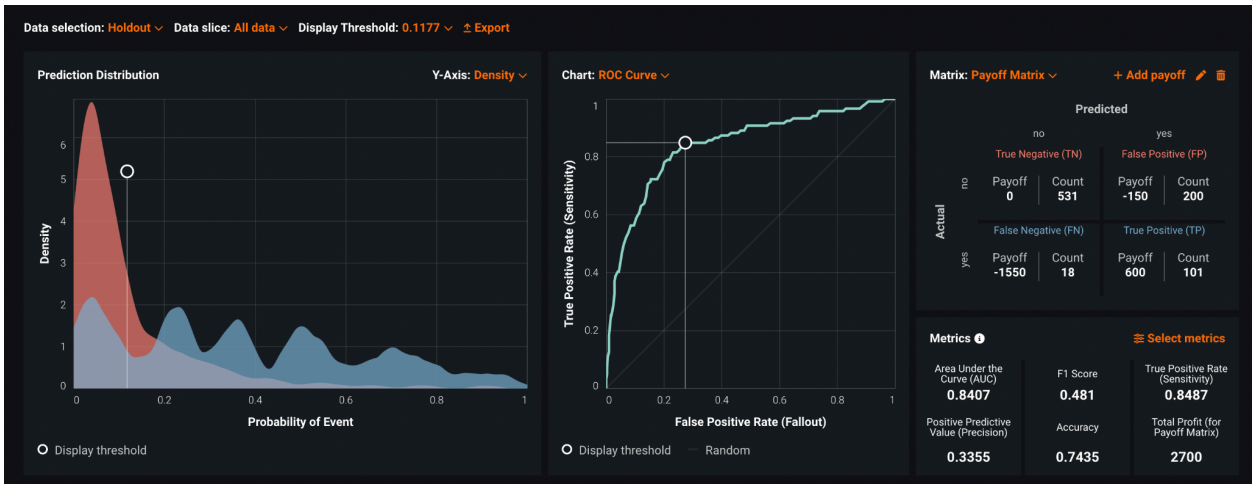
eXtreme Gradient Boosted Trees Classifier (XGBoost): This model's power in handling complex datasets with mixed types of features, like those found in telecom data, allows it to detect intricate patterns that might indicate churn risk. Its robustness and ability to model non-linear decision boundaries are particularly relevant for telco datasets that contain a mix of demographic, usage, and service-related features.

RandomForest Classifier (Entropy): A RandomForest can efficiently model the random and often non-linear nature of customer behavior in the telecommunications industry. Its ability to

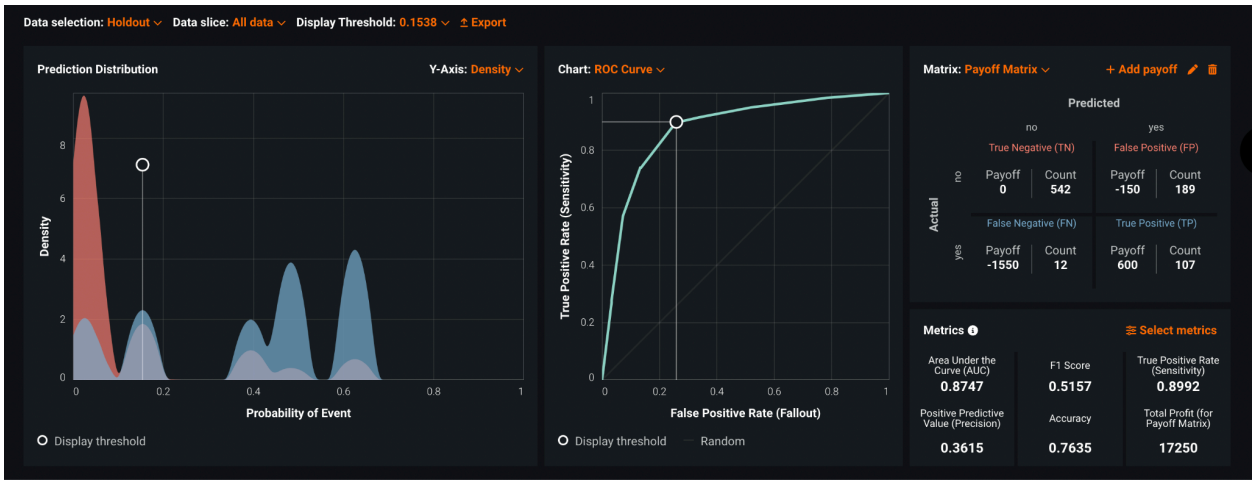
handle a large number of input variables is relevant for dealing with the complex feature sets often present in telecom data, such as call minutes, plan details, and customer service interactions.



Logistic Regression: This model provides a straightforward approach that can yield valuable insights into which features have the strongest influence on churn. Its output can be interpreted in terms of the probability of a customer churning, which is highly valuable for business stakeholders looking for direct, actionable insights.



Decision Tree Classifier (Gini): Decision trees can be particularly useful for identifying key decision points that lead to customer churn, such as thresholds in service usage or satisfaction levels. They are easy to understand and can be used to communicate the findings to non-technical stakeholders, aiding in strategic decision-making.



	eXtreme Gradient Boosted Trees Classifier (XGBoost)	RandomForest Classifier (Entropy)	Logistic Regression	Decision Tree Classifier (Gini)
Recall	0.8319	0.8487	0.8487	0.8992
Precision	0.7226	0.6516	0.3355	0.3615
F1	0.7734	0.7372	0.481	0.5157
Accuracy	0.9318	0.9153	0.7435	0.7635
Error	0.0682	0.0847	0.2565	0.2365
ROC AUC	0.9025	0.9066	0.8407	0.8747
Maximum Payoff	$22,700	$24,600	$2,700	$17,250
LogLoss	0.1729	0.1762	0.3518	0.4250
Cross Validation	0.1645	0.1654	0.3302	0.4142
Holdout	0.1798	0.1846	0.2998	0.3432
Threshold	0.1285	0.1415	0.1177	0.1538

Explaining Metrics

- ***Recall (True Positive Rate):*** Higher recall means fewer false negatives, crucial when the cost of overlooking a potential churn is significant.
- ***Precision:*** Indicates the accuracy of positive predictions, with higher precision reducing the number of false positives, essential for minimizing unnecessary marketing efforts.

- ***F1 Score:*** Balances precision and recall, useful for finding an optimal blend of both metrics.
- ***Accuracy:*** Represents the proportion of true results (both TP and TN) among the total number of cases examined.
- ***ROC AUC:*** Measures the model's ability to distinguish between classes. Higher values indicate better performance.
- ***Maximum Payoff:*** Estimates the financial benefit of the model, considering the costs and benefits of each prediction outcome. A higher maximum payoff suggests a greater financial benefit.
- ***LogLoss:*** Provides a measure of accuracy for a classifier, with lower values indicating better performance. It penalizes false classifications.
- ***Cross Validation:*** Averages the model's effectiveness over multiple data splits to ensure reliability.
- ***Holdout:*** Evaluates the model on a separate dataset not used in training to test its generalization capability.

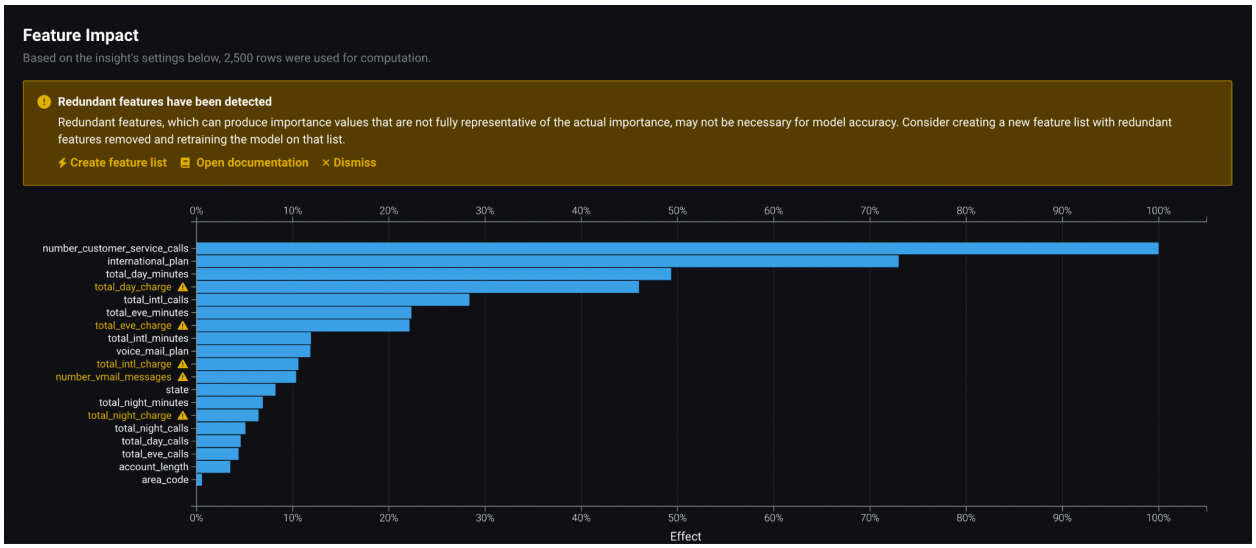
Performance Summary:

- XGBoost shows robust performance with the lowest LogLoss and strong Cross Validation and Holdout scores, indicating high reliability and accuracy in its predictions.
- RandomForest Classifier stands out with the highest Maximum Payoff of $24,600, signifying its superior financial benefit. It also exhibits solid ROC AUC and Accuracy, balancing churn prediction accuracy with the costs of misclassification effectively.
- Logistic Regression and Decision Tree Classifier display higher recall but lower precision, which could result in more false positives and potentially lower financial efficiency.

Best Model

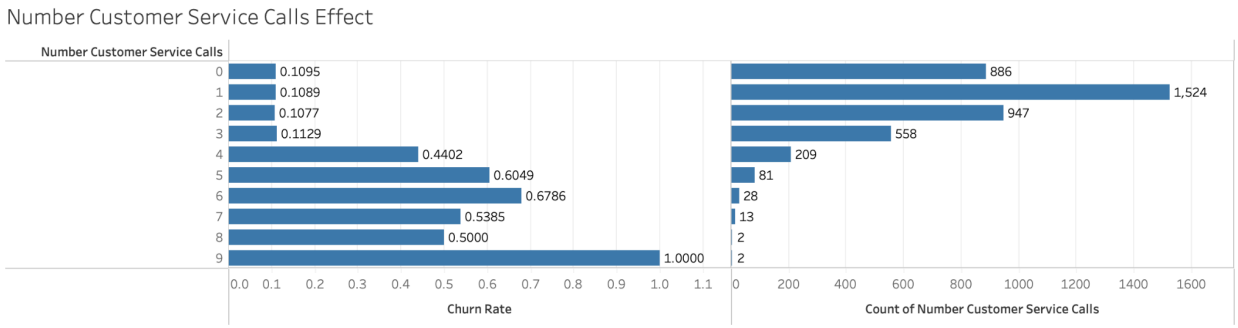
- Given the analysis's focus on optimizing marketing campaign efficiency and resource allocation, the RandomForest Classifier emerges as the most financially advantageous model due to its highest Maximum Payoff. This metric underscores the economic value of the model, weighing the benefits of correct predictions against the costs of misclassifications. While the RandomForest Classifier demonstrates exceptional ability in distinguishing between customers likely to churn and those who will not, thereby maximizing economic returns for the company, XGBoost also presents a strong case with the best LogLoss, indicating high predictive accuracy and reliability.
- The Logistic Regression and Decision Tree Classifier, despite their usefulness in certain contexts, offer lower financial efficiency due to their tendency towards higher false positives. Therefore, in the context of maximizing financial outcomes and marketing efficiency, the RandomForest Classifier is recommended as the primary model for deployment.

Q4. Visualize the effects of the top 4 predictors of customer churn. Summarize the effects in 1-2 sentences.

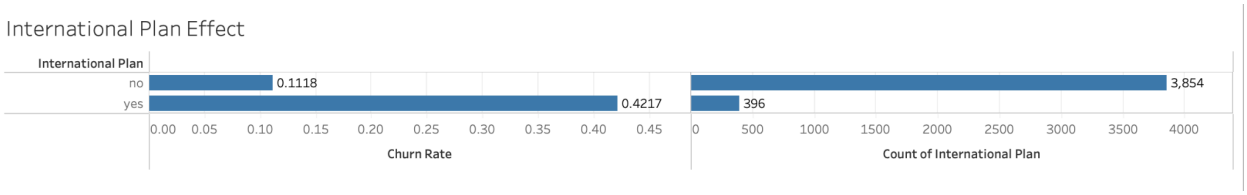


Using Datarobot, we are able to see the feature impact, which includes all the features present in the "telco_churn" data set, as it was concluded during the exploratory data analysis that none of them were 100% redundant. As such, for the RandomForest Classifier model, the the most financially beneficial model with the highest Maximum Payoff, the top four predictors are:

- ***Number Customer Service Calls:*** A higher number of calls to customer service may indicate issues or dissatisfaction with the service, which can lead to churn.

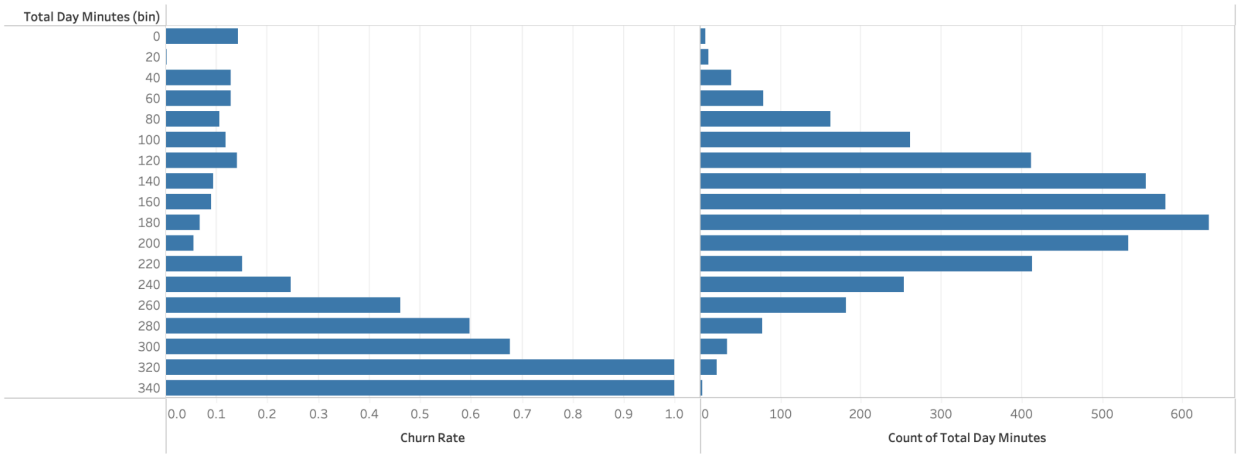


- ***International Plan:*** The presence of an international plan could influence churn, suggesting that those with international plans might be more likely to churn, possibly due to the cost or service quality associated with international calling.



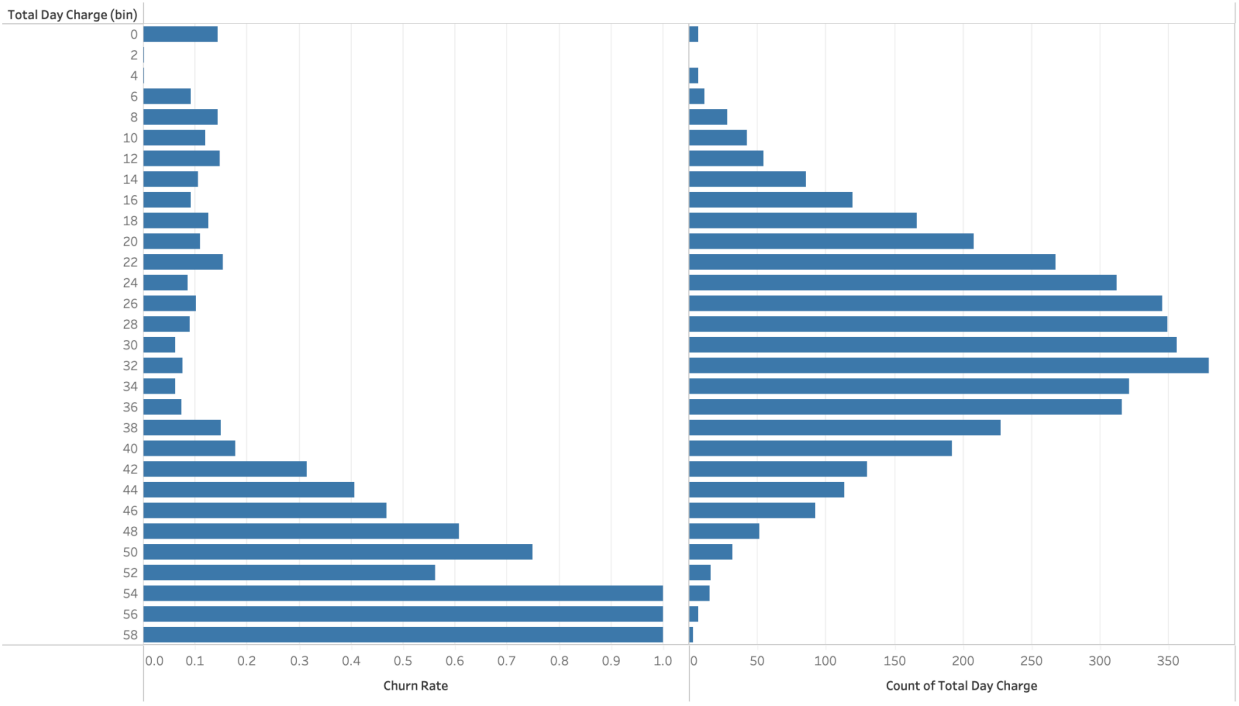
- ***Total Day Minutes:*** Churn rate generally increases as the total day minutes increase, with the highest churn rates occurring in the higher usage bins. This suggests that customers with higher day minutes usage are more likely to churn, possibly due to the higher costs incurred during peak hours.

Total Day Minutes Effect



- ***Total Day Charge:*** This pattern suggests that higher total day charges could be a significant factor in customer churn, implying that customers experiencing higher charges may be more likely to leave the service. This could be due to perceived value, dissatisfaction with cost, or better competing offers.

Total Day Charge Effect



Note: For this visualization, the target (churn rate) is on the y-axis in the top panel, and the count of customer calls is on the y-axis. This visualization approach helps to clearly illustrate the relationship between the key predictors identified by the RandomForest Classifier and the outcome of interest, which in this case is customer churn. By focusing on the churn rate and the count of customer calls, stakeholders can gain insightful perspectives on how these predictors influence the likelihood of churn, facilitating more informed decision-making to enhance retention strategies and ultimately improve the financial performance of the company.

Q5. For each of the top 4 predictors, formulate actionable recommendations based on the observed effects. If the observed effect cannot be reasonably made actionable, please state so.

- ***Number of Customer Service Calls:*** A high frequency of such calls is typically a signal of customer issues or dissatisfaction with the service. To mitigate this, the company should establish a system that triggers a follow-up process after a customer reaches a threshold number of calls, ensuring their concerns are being resolved. Moreover, a detailed analysis of the content of these calls could uncover common complaints or issues, which can then be systematically addressed to improve product or service quality, potentially reducing the volume of calls and subsequently the churn rate.
- ***International Plan:*** The observed higher churn rates among customers who subscribe to an international plan suggest that the plan may be misaligned with customer expectations or market standards. The company would benefit from re-evaluating the pricing and features of these plans, ensuring they offer competitive value. Additionally, actively seeking feedback from subscribers of these plans could provide insights into their specific needs and inform the development of more personalized or tiered international plan options.
- ***Total Day Minutes:*** The data suggests that customers with higher usage during the day are more prone to churn, likely due to the higher charges incurred. In response, the company should consider revising their pricing strategy to offer more economical plans that allow high-usage customers to access more minutes at a lower cost. Encouraging off-peak call times with discounted rates could also alleviate high usage charges. Furthermore, providing additional benefits or services to heavy users could enhance their perceived value of the service, fostering customer loyalty.
- ***Total Day Charges:*** The pattern observed indicates that higher charges may lead to increased churn. To address this, the company could implement a system that alerts customers as they near high-charge levels or offer a cap on daily charges to prevent unexpected high bills. Ensuring transparency in billing and providing a clear breakdown of charges could help customers understand their expenses better and reduce the risk of bill shock. Offering loyalty discounts or rewards for consistent usage may also serve as an incentive for customers to remain with the company despite higher usage charges.