Predicting Churn in a Telecom Company

Team Members:

Jasvitha Vatsavaya

Manasa Ramaka

Sai Abhinav Mullapudi

Vaikhari Tushar Kadam

Contents

Executive Summary:

The project, titled "Predicting Churn in a Telecom Company," investigates customer churn within a fictitious telecom company serving 7,043 customers in California during the third quarter. Churn, the process where customers discontinue their services, is a critical issue for the telecom industry, directly impacting revenue and customer retention strategies. Leveraging the Telco Customer Churn dataset, this study aimed to uncover patterns and factors associated with churn and develop predictive models to help businesses proactively address this challenge.

The project followed a systematic approach involving data exploration, feature engineering, and modeling. Data exploration provided insights into customer demographics, service preferences, and account behaviors, revealing key trends in churn. Feature engineering transformed raw data into meaningful variables, enhancing model performance. Several models learned in the course were applied and compared, allowing for a thorough evaluation of their suitability for the dataset. This comparison helped identify the best-performing model to move forward with, ensuring that the most effective predictive tool was selected for tackling the problem.

This project showcases the practical application of classroom knowledge to address real-world challenges, emphasizing the importance of combining analytical techniques with domain knowledge. The insights generated provide actionable strategies for telecom companies to reduce churn, improve customer satisfaction, and foster long-term loyalty. By integrating data exploration, feature engineering, and model evaluation, the study highlights the value of data-driven decision-making in building effective business strategies in the competitive telecom sector.

Data Dictionary:

customerID: It is a unique identifier to find a particular customer in the database. (character variable)

Gender: Specifies the gender of the customer. (character variable)

SeniorCitizen: Specifies if the customer is a senior citizen or not. (character variable)

Partner: Specifies if the particular customer has a partner or not. (character variable)

Dependents: Specifies if the particular customer has a dependent or not. (character variable)

Tenure: The total amount of months the customer has been with the company. (numeric continuous)

PhoneService: Does the customer have a phone service or not. (character variable)

MultipleLines: Whether the customer has multiple lines or not. (character variable)

InternetService: Customer's internet connection type. (character variable)

OnlineSecurity: If the customer has opted for the online security feature with the company. (character variable)

OnlineBackup: If the customer has opted for the online backup feature with the company. (character variable)

DeviceProtection: If the customer has opted for the device protection feature with the company. (character variable)

TechSupport: If the customer has opted for the technical support feature with the company. (character variable)

StreamingTV: If the customer has opted for the online tv streaming services with the company. (character variable)

StreamingMovies: If the customer has opted for the online movies streaming services with the company. (character variable)

Contract: Current contract length to which the customer is tied. (numeric continuous)

PaperlessBilling: If the customer has opted for a paperless billing option. (character variable)

PaymentMethod: The customer's method of transaction. (character variable)

MonthlyCharges: The customer's current monthly charges. (numeric continuous)

TotalCharges: The total cumulative charges the customer has paid in his tenure as a customer to the company. (numeric continuous)

Churn: The customers who left in the last quarter. (character variable)

Feature Engineering:

1. **Churn** is the target variable for our project.
2. We have converted **gender** (male=1, female=0), **seniorcitizen** (yes=1, no=0), **partner** (yes=1, no=0), **dependent** (yes=1, no=0), **phoneservice** (yes=1, no=0), **multiplelines** (yes=1, no phone service/no=0), **internetservice** (yes=1, no phone service/no=0), **onlinesecurity** (yes=1, no phone service/no=0), **onlinebackup** (yes=1, no phone

service/no=0), **deviceprotection** (yes=1, no phone service/no=0), **techsupport** (yes=1, no phone service/no=0), **streamingtv** (yes=1, no phone service/no=0), **streamingmovie** (yes=1, no phone service/no=0), **paperlessbilling** (yes=1, no=0) and **churn** (yes=1, no=0) to numeric continuous variables.

3. We have created two new columns:

   a. **Total Services:** It would tell us the number of services the customer has availed from the company.

   Total Services= onlinebackup + deviceprotection + techsupport + streamingtv + streamingmovie .

   We have not added internetservice in total service as we felt that individually internetservice could play a key role in the predictive model. We deleted the other features from the model retaining total services and internetservice.

   b. **MonthlyChargeDiff:** It would tell us if the customer has changed his plan or continues to use the same plan in his tenure as a customer. It would also tell if the customer has downgraded his plan or upgraded his plan.

   MonthlyChargeDiff= MonthlyCharge - Avg. MonthlyCharge

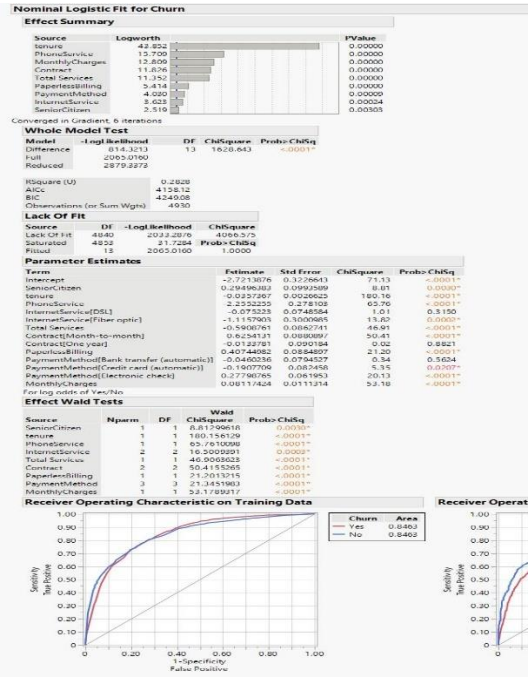
   Here avg MonthlyCharge= TotalCharges/Tenure.

   If the MonthlyChargeDiff >0 we can know he might have upgraded his plan if not he might have downgraded his plan. If the MonthlyChargeDiff=0 the customer might have not changed his plan at all. We have deleted the TotalCharges from the model.

Modeling:

We have used random seed 888 wherever necessary. For the validation set we have divided the data into 70% training and 30% validation data.

Logistic Regression:



This is the final model that we got in Logistic Regression.

Approach:

After fitting the base model, We checked the p-values and re-fitted the model. It took a total of 6 iterations (including base model) to get to the final model. In the process we removed gender, partner, dependents, multiple lines, monthlychargeDiff.

Takeaways:

1. The best predictor here is **tenure** with a logworth value of 43.852. [1]

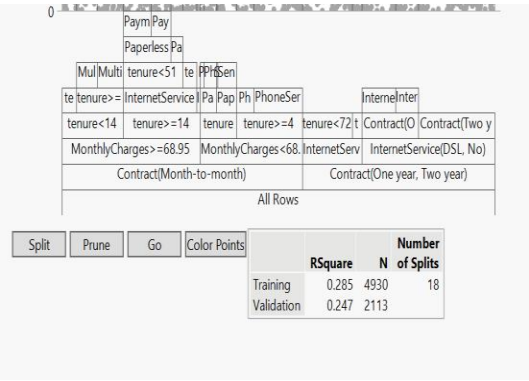
2. The model is statistically significant, with a **Chi-Square value of 1628.643** and a p-value of <0.0001. This tells that the predictors collectively explain a significant portion of the variance in churn.

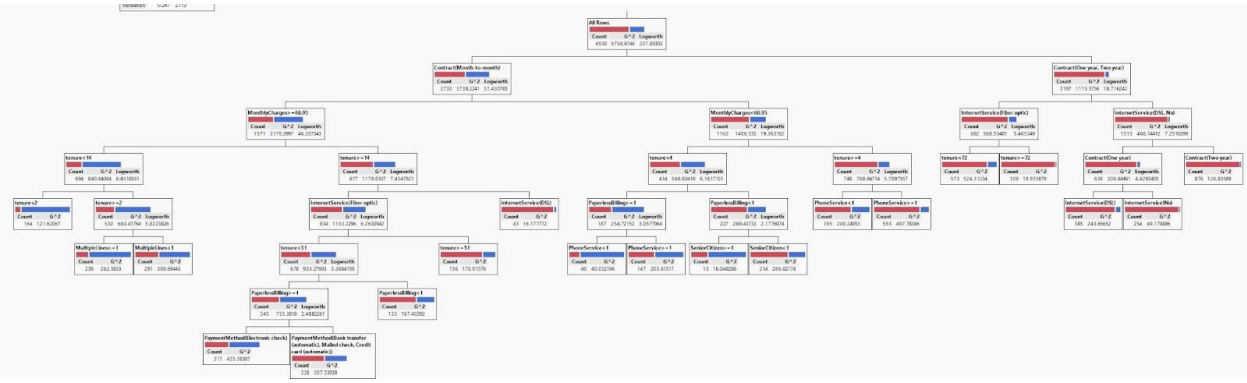
3. The **lack-of-fit test p-value is 1.0000**, meaning there is no significant lack of fit. The model converged efficiently (in just 6 iterations), showing it is well-calibrated without overfitting.

4. The model has an **AUC** value of 0.8463 and 0.8375 for training and validation data respectively. The high AUC values indicates that the model is effective in predicting the churn.

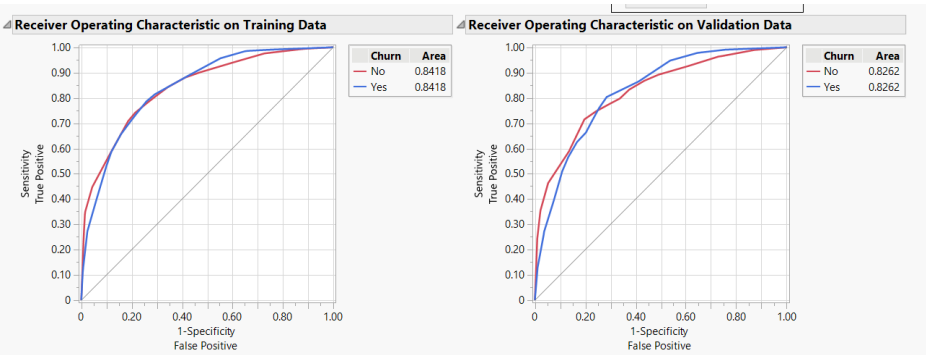
Decision Tree:

After fitting the Decision Tree model, we identified 18 splits driven by key predictors such as tenure, monthly charges, contract type, and internet service type. Further splits were avoided due to diminishing R-Squared gains and the risk of overfitting. Finalizing the model at 18 splits ensured it remained interpretable, efficient, and generalizable.





ROC Curve and AUC:



Takeaways:

1. **Contract Type Dominates:** The most significant factor is the split on month-to-month contracts, highlighting that these customers are at the greatest risk of churn.

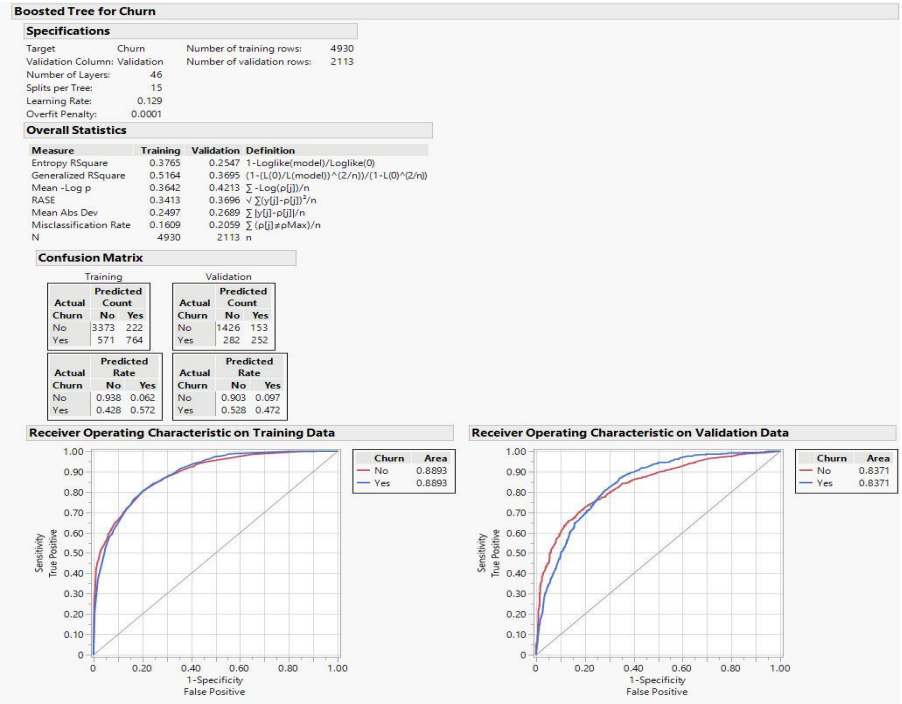
2. **Tree Complexity:** The decision tree has 18 splits, showing it captures meaningful patterns while remaining interpretable.

3. **Tenure as a Key Factor:** Low tenure (less than 4 months) appears frequently in splits, indicating that newer customers are more likely to churn.

4. The model achieves an AUC of 0.8418 on training data and 0.8262 on validation data, demonstrating strong predictive performance and good generalization to new data. The

slight drop in AUC from training to validation indicates minimal overfitting, highlighting the model's robustness.

Boosted Tree:



Boosted Tree for Churn

Specifications

Target	Churn	Number of training rows:	4930
Validation Column:	Validation	Number of validation rows:	2113
Number of Layers:	46		
Splits per Tree:	15		
Learning Rate:	0.129		
Overfit Penalty:	0.0001		

Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.3765	0.2547	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5164	0.3695	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3642	0.4213	∑ -Log(p[j])/n
RASE	0.3413	0.3696	√ ∑(y[j]-p[j])²/n
Mean Abs Dev	0.2497	0.2689	∑ |y[j]-p[j]|/n
Misclassification Rate	0.1609	0.2059	∑ (p[j]≠pMax)/n
N	4930	2113	n

Confusion Matrix

Training

Actual Churn	Predicted Count No	Predicted Count Yes
No	3373	222
Yes	571	764

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.938	0.062
Yes	0.428	0.572

Validation

Actual Churn	Predicted Count No	Predicted Count Yes
No	1426	153
Yes	282	252

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.903	0.097
Yes	0.528	0.472

Receiver Operating Characteristic on Training Data

Receiver Operating Characteristic on Validation Data

Takeaways:

1. The model achieves a Generalized R^2 of 0.5164 for training and 0.3695 for validation, indicating good explanatory power with reasonable generalization.
2. The model's 46 layers and 15 splits per tree strike a balance between complexity and predictive accuracy.
3. The validation results show the model effectively identifies churners but could improve recall by reducing false negatives.

4. The model achieved an AUC of 0.8893 on the training data and 0.8371 on the validation data, demonstrating strong discriminatory power, good generalization, and consistent performance across datasets with minimal overfitting.

Neural Network:

Model NTanH(3)

Training

Churn

Measures	Value
Generalized RSquare	0.4226228
Entropy RSquare	0.2946584
RASE	0.365295
Mean Abs Dev	0.2673712
Misclassification Rate	0.1908722
-LogLikelihood	2030.9163
Sum Freq	4930

Confusion Matrix

Actual Churn	Predicted Count No	Yes
No	3268	327
Yes	614	721

Confusion Rates

Actual Churn	Predicted Rate No	Yes
No	0.909	0.091
Yes	0.460	0.540

Validation

Churn

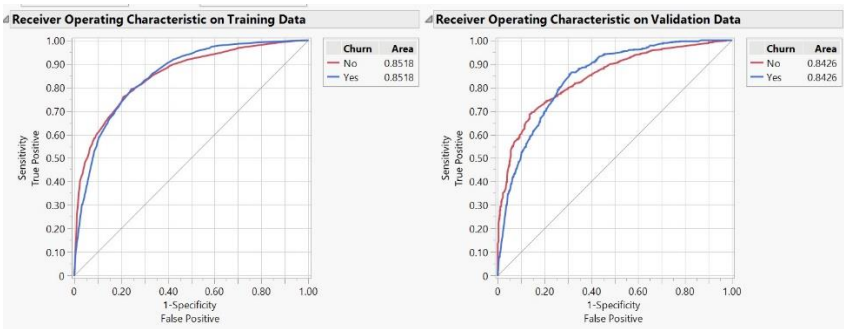
Measures	Value
Generalized RSquare	0.3894869
Entropy RSquare	0.2708101
RASE	0.3669711
Mean Abs Dev	0.2659105
Misclassification Rate	0.2001893
-LogLikelihood	871.00966
Sum Freq	2113

Confusion Matrix

Actual Churn	Predicted Count No	Yes
No	1425	154
Yes	269	265

Confusion Rates

Actual Churn	Predicted Rate No	Yes
No	0.902	0.098
Yes	0.504	0.496



Approach:

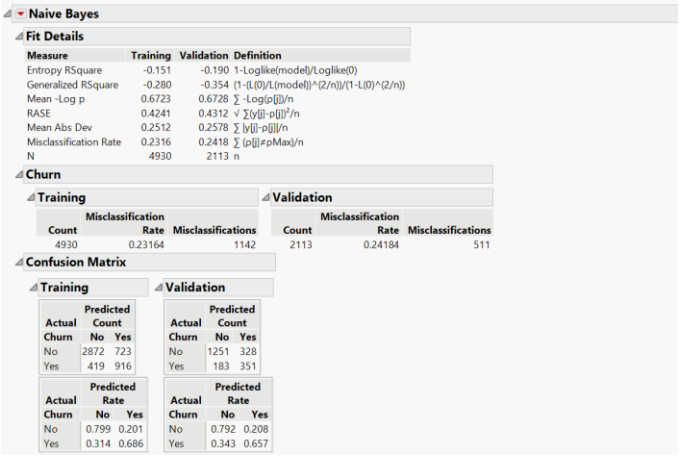
We evaluated three neural network configurations using AUC on validation data. The first model, with one hidden layer and three nodes using the TanH activation function, achieved the highest AUC of 0.8426. Increasing complexity to one hidden layer with six nodes (AUC: 0.8406) or two

hidden layers with three nodes each (AUC: 0.8354) slightly reduced performance. The initial configuration was selected as the final model for its superior fit. [2]

Takeaways:

1. **Generalized R^2**: The model achieves a Generalized R^2 of 0.4226 (training) and 0.3895 (validation), indicating it explains a substantial portion of the variance and generalizes well.
2. **Misclassification Rates**: With misclassification rates of 0.1908 (training) and 0.2002 (validation), the model demonstrates strong accuracy and minimal overfitting.
3. **ROC AUC**: The AUC values of 0.8518 (training) and 0.8426 (validation) confirm excellent performance in distinguishing churners from non-churners, supported by ROC curves showing high sensitivity with minimal false positives.
4. **Model Speed**: The model runs efficiently, ensuring quick performance without compromising accuracy.

Naive Bayes:



Naive Bayes

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	-0.151	-0.190	1-Loglike(model)/Loglike(0)
Generalized RSquare	-0.280	-0.354	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.6723	0.6728	∑ -Log(p[j])/n
RASE	0.4241	0.4312	√ ∑(y[j]-p[j])²/n
Mean Abs Dev	0.2512	0.2578	∑ |y[j]-p[j]|/n
Misclassification Rate	0.2316	0.2418	∑ (p[j]≠pMax)/n
N	4930	2113	n

Churn

Training

Count	Misclassification Rate	Misclassifications
4930	0.23164	1142

Validation

Count	Misclassification Rate	Misclassifications
2113	0.24184	511

Confusion Matrix

Training

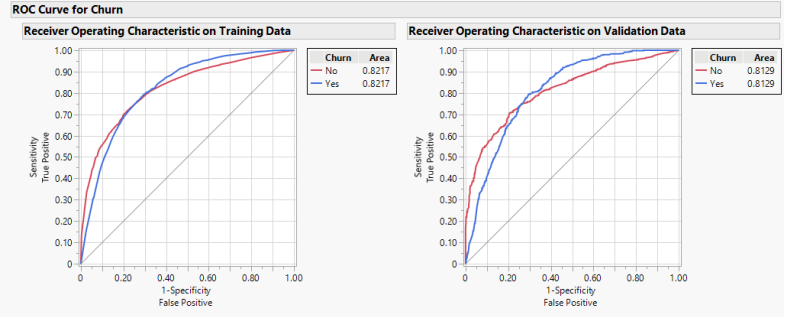
Actual Churn	Predicted Count No	Predicted Count Yes
No	2872	723
Yes	419	916

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.799	0.201
Yes	0.314	0.686

Validation

Actual Churn	Predicted Count No	Predicted Count Yes
No	1251	328
Yes	183	351

Actual Churn	Predicted Rate No	Predicted Rate Yes
No	0.792	0.208
Yes	0.343	0.657



Takeaways:

- The model demonstrated low misclassification rates of 23.16% on training and 24.18% on validation, reflecting strong generalizability across datasets.
- The model accurately predicted 79.9% of non-churners and 68.6% of churners in the training data, with validation results slightly lower but consistent, at 79.2% and 65.7%, respectively.
- The Naive Bayes model performed well, achieving AUC values of 0.8217 (training) and 0.8129 (validation), indicating reliable discriminatory power between churners and non-churners reflecting minimal overfitting and strong generalization.

Model Comparison:

Model Used	**AUC of Training Set Achieved**	**AUC of Validation Set Achieved**
Logistic Regression	0.8463	0.8375
Decision Tree	0.8418	0.8262
Boosted Tree	0.8893	0.8371
Neural Networks	0.8518	0.8426

Naive Bayes	0.8217	0.8129

Based on the AUC values obtained from the validation sets, we conclude that the neural network model demonstrates the best performance and is the most suitable choice for our Project. It provides a balance between complexity and generalization, ensuring reliable performance on unseen data. While other models like Boosted Tree also show strong performance, Neural Networks' validation metrics and robust generalization make it the most suitable choice for real-world deployment.

Business Understanding:

Customer churn is one of the biggest difficulties that the telecom industry encounters. High churn rates cause revenue losses, higher client acquisition expenses, and low profitability. Retaining existing customers is significantly more cost effective than recruiting new ones, thus churn prevention is a primary goal. The Telecom Customer turnover dataset offers a comprehensive perspective of customer behavior, service consumption, and account information, providing significant insights into the elements that contribute to turnover. Understanding these characteristics might help telecom businesses design customer retention strategies and avoid churn-related losses.

The business problem of this project aims to identify the customers being at the high risk of churn and understand the reasons behind their dissatisfaction. Key factors like contract types, monthly charges and services can significantly contribute to the dissatisfaction of the customers thus making them influenced to leave. Analyzing these patterns and building a predictive model can certainly help in retaining the existing customers and bringing in new customers. For

instance, offering discounts to customers monthly or customizing their plan accordingly can significantly improve the satisfaction rate.

Solving this problem is vital for maintaining a competitive edge in the highly saturated telecom market. A data-driven approach to churn prediction enables businesses to allocate resources efficiently and enhance customer satisfaction. Retention efforts based on these insights can lead to a loyal customer base, increased revenue, and a stronger reputation in the market. This project underscores the importance of leveraging data analytics to address critical business challenges and drive growth.

Evaluation

In the telecommunication industry, where churn rates range from 25-35%, predicting customer churn is essential to mitigate revenue loss and maintain brand value. Key metrics like AUC assess the model's ability to differentiate churners from non-churners, while misclassification rates and the confusion matrix focus on minimizing false negatives and identifying true positives, aiding in the design of effective retention strategies. Generalized R-Squared (R^2) values further validate how well independent variables explain churn variability, ensuring the model's reliability.

The predictions can drive targeted retention campaigns such as tailored offers, loyalty programs, and engagement strategies, allowing companies to optimize budgets and maximize returns. For example, retaining 20% of churners with a $50 cost per customer and an average lifetime value of $5,000 can significantly boost revenue. When ROI estimation is difficult due to customer variability, sensitivity analysis can provide revenue impact projections. Alternatively, actionable

insights from the model can inform broader strategies like improving service, refining pricing, or introducing new bundles, indirectly reducing churn rates.

Deployment

The churn prediction model, built using JMP Predictive Modeling, will be deployed as an integral part of the company's CRM system to identify and address high-risk customers. The deployment will focus on three key areas:

- Customer Segmentation: Using the model's predictions, customers will be segmented based on their likelihood to churn, enabling targeted retention strategies.
- Actionable Insights: Dashboards will display real-time predictions and key metrics, giving marketing and customer service teams the tools to offer personalized loyalty programs or tailored discounts.
- Trend Monitoring: By tracking churn patterns, the business can proactively adjust strategies, ensuring retention efforts stay aligned with customer needs.

For scalability and flexibility, Python could be leveraged to rebuild the model, allowing integration with automation pipelines and APIs for seamless deployment. Python's libraries like Scikit-learn and Flask will enable dynamic retraining and efficient delivery of predictions to stakeholders.

Issues to Consider

- Data Accuracy: The quality of input data directly impacts predictions. Regular audits will be critical to maintaining accuracy and reliability.

- Stakeholder Adoption: The business teams need clear training on how to interpret model outputs and act on insights, ensuring the predictions translate into actionable decisions.

Ethical Considerations

- Bias Auditing: The model must be periodically reviewed to ensure fairness, avoiding any unintentional discrimination based on customer demographics.
- Privacy Concerns: Adhering to GDPR or similar data privacy regulations is non-negotiable. Customer data must be anonymized where possible to mitigate risks.
- Transparency: Clear explanations of how predictions are made will help build trust with internal teams and external stakeholders.

Risk Mitigation

- Overfitting and Generalization: Regular evaluation and retraining with updated datasets will ensure the model stays relevant and accurate.
- Stakeholder Resistance: Early involvement of end-users in testing and iterative development will encourage adoption and trust in the model.

Deploying the model through JMP ensures immediate usability, while exploring a transition to Python sets the foundation for long-term scalability and automation. This dual approach ensures the company can make data-driven decisions, reduce churn, and ultimately drive sustainable business growth.

References

- https://www.kaggle.com/code/mehmetisik/telecom-churn-prediction-learning-ml-models/notebook
- https://link.springer.com/chapter/10.1007/978-3-319-62416-7_28
- https://ieeexplore.ieee.org/abstract/document/6693977
- https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6
- https://ieeexplore.ieee.org/abstract/document/7528311
- Comprehensive Report: Telecom Customer Churn Analysis and Recommendations | by Henry Chukwunwike Morgan-Dibie | Medium
- https://www.nelsonmullins.com/insights/alerts/additional_nelson_mullins_alerts/all/fcc-cpni-certification-and-privacy-rules-update
- https://bja.ojp.gov/program/it/privacy-civil-liberties/authorities/statutes/1285

Appendix:

1. **Odds Ratio:**

Nominal Logistic Fit for Churn

Converged in Gradient, 6 iterations

Odds Ratios

For Churn odds of Yes versus No

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
SeniorCitizen	1.343078	1.105279	1.631801	0.7445585
tenure	0.964894	0.959832	0.969904	1.036383
PhoneService	0.10485	0.060632	0.180417	9.5374435
Total Services	0.553842	0.467341	0.655464	1.8055697
PaperlessBilling	1.50298	1.264036	1.78832	0.6653448
MonthlyCharges	1.08456	1.061247	1.108592	0.922033

Range Odds Ratios

Per change in regressor over entire range

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
SeniorCitizen	1.343078	1.105279	1.631801	0.7445585
tenure	0.076303	0.052245	0.110786	13.105661
PhoneService	0.10485	0.060632	0.180417	9.5374435
Total Services	0.028861	0.010419	0.079304	34.648584
PaperlessBilling	1.50298	1.264036	1.78832	0.6653448
MonthlyCharges	3491.235	393.1444	31590.28	0.0002864

Odds Ratios for InternetService

Level1	/Level2	Odds Ratio	Prob>Chisq	95% Confidence Interval (Wald) Lower	Upper
Fiber optic	DSL	0.353254	0.0012*	0.188311	0.662674
No	DSL	3.547476	<.0001*	1.922151	6.547136
No	Fiber optic	10.04227	<.0001*	3.143767	32.07849
DSL	Fiber optic	2.830823	0.0012*	1.509038	5.310376
DSL	No	0.281891	<.0001*	0.152739	0.52025
Fiber optic	No	0.099579	<.0001*	0.031174	0.31809

Odds Ratios for Contract

Level1	/Level2	Odds Ratio	Prob>Chisq	95% Confidence Interval (Wald) Lower	Upper
One year	Month-to-month	0.52793	<.0001*	0.411982	0.676511
Two year	Month-to-month	0.290124	<.0001*	0.197066	0.427125
Two year	One year	0.549549	0.0028*	0.371225	0.813536
Month-to-month	One year	1.89419	<.0001*	1.478172	2.427292
Month-to-month	Two year	3.446806	<.0001*	2.341238	5.074441
One year	Two year	1.819673	0.0028*	1.229202	2.693787

Odds Ratios for PaymentMethod

Level1	/Level2	Odds Ratio	Prob>Chisq	95% Confidence Interval (Wald) Lower	Upper
Credit card (automatic)	Bank transfer (automatic)	0.865241	0.2853	0.663488	1.128342
Electronic check	Bank transfer (automatic)	1.382663	0.0037*	1.110659	1.721281
Mailed check	Bank transfer (automatic)	1.004842	0.9712	0.773251	1.305795
Electronic check	Credit card (automatic)	1.598009	<.0001*	1.272529	2.006738
Mailed check	Credit card (automatic)	1.161344	0.2753	0.887655	1.519418
Bank transfer (automatic)	Credit card (automatic)	1.155747	0.2853	0.886256	1.507185
Mailed check	Electronic check	0.726744	0.0048*	0.582134	0.907278
Bank transfer (automatic)	Electronic check	0.723242	0.0037*	0.580963	0.900366
Credit card (automatic)	Electronic check	0.625779	<.0001*	0.498321	0.785837
Bank transfer (automatic)	Mailed check	0.995181	0.9712	0.765817	1.293241
Credit card (automatic)	Mailed check	0.861071	0.2753	0.658147	1.126564
Electronic check	Mailed check	1.376	0.0048*	1.102198	1.717818

Normal approximations used for ratio confidence limits effects: InternetService Contract PaymentMethod

Tests and confidence intervals on odds ratios are Wald based.

- Senior citizens have 34.3% higher odds of churn compared to non-senior customers (Odds Ratio = 1.3431).

- Longer tenure significantly reduces churn risk, with each additional month lowering odds by 3.5% (Odds Ratio = 0.9648).

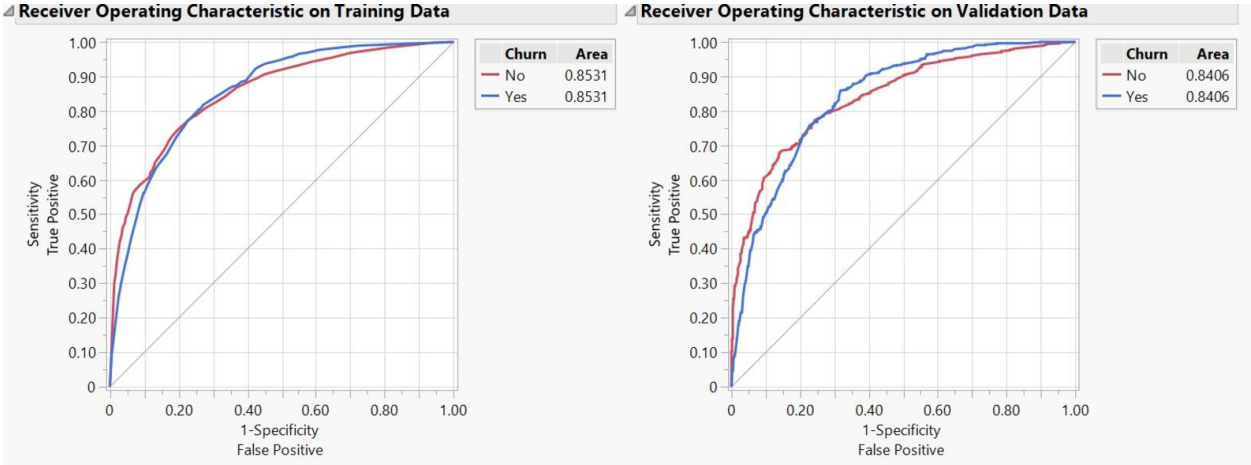
- Higher monthly charges increase churn odds by 8.46% for every $1 increase (Odds Ratio = 1.0846).

- Customers on month-to-month contracts are 3.4 times more likely to churn than those on two-year contracts (Odds Ratio = 3.4486).

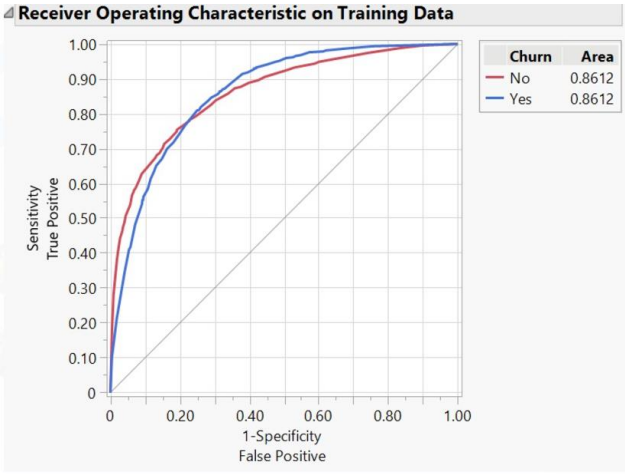
- Paying via automatic credit card reduces churn risk compared to bank transfers (Odds Ratio = 0.8652).

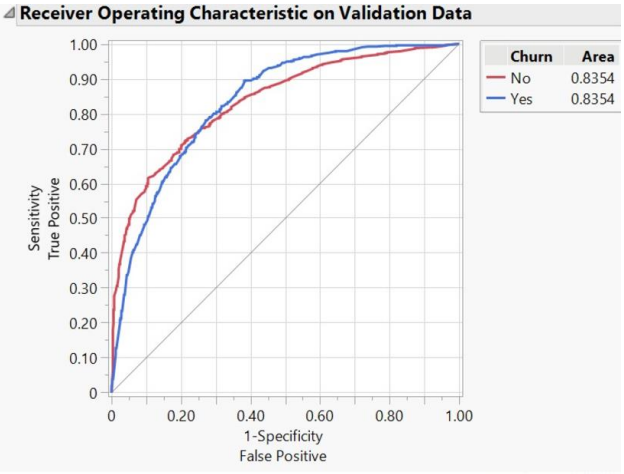
2. In this evaluation, three neural network configurations were tested using AUC on validation data.

   - The first model, with one hidden layer and three nodes using the TanH activation function, achieved the highest AUC of 0.8426, indicating the best performance.
   - The second model, with one hidden layer and six nodes, resulted in a slight drop in AUC to 0.8406, suggesting the added complexity didn't improve performance.



   - The third model, with two hidden layers and three nodes each, had the lowest AUC of 0.8354. These results indicated that the first model provided the best balance between complexity and performance, making it the final choice.


Receiver Operating Characteristic on Training Data
Sensitivity
True Positive
1-Specificity
False Positive
Churn
Area
No
0.8612
Yes
0.8612


Receiver Operating Characteristic on Validation Data
Sensitivity
True Positive
1-Specificity
False Positive
Churn
Area
No
0.8354
Yes
0.8354