SRM UNIVERSITY - DEPARTMENT OF COMPUTATIONAL INTELLIGENCE

SCHOOL OF COMPUTING

CASE STUDY REPORT

Course Code: 21AIC401T
Course Name: Inferential Statistics and Predictive Analytics
Assignment Type: Case Study-Based Modeling Project
Title: *Customer Churn Prediction – Model Development, Validation, and Deployment*
Student Name: Pearl Rubyth Thomas R
Register No.: RA2212701010021
Date:10th November 2025

---

1. Introduction and Data Preparation

1.1 Objective

Customer churn represents one of the most pressing challenges for subscription-based and telecom industries. Churn occurs when an existing customer stops using the company's services, leading to revenue loss and increased acquisition costs for replacements. Predicting which customers are likely to churn enables proactive retention strategies—such as loyalty discounts, improved support, or service bundling—thus improving profitability and sustainability.

The aim of this project is to develop and validate a predictive model capable of accurately identifying customers at risk of churn. The modeling process combines **statistical inference** (Generalized Additive Models) with **rule-based learning** (CHAID Decision Trees) to achieve both interpretability and accuracy.

The objectives are:

- Download and prepare a real-world telecom dataset.
- Clean, analyze, and visualize customer behavior.
- Build and compare statistical and rule-based models (CHAID and GAM).
- Deploy the best-performing model for real-time prediction.

1.2 Dataset Description

Dataset Source: Kaggle - *Telco Customer Churn* (blastchar/telco-customer-churn)
Total Records: 7,043
Attributes: 21 columns (15 categorical, 6 numerical)
Target Variable: Churn (1 = customer left, 0 = retained)

Type	Feature Examples	Description
Numerical	tenure, MonthlyCharges, TotalCharges	Customer duration and billing
Categorical	Contract, InternetService, PaymentMethod, OnlineSecurity, etc.	Subscription and service details
Target	Churn	Customer retention label

1.3 Data Cleaning

Steps taken:

- Converted TotalCharges from string to numeric.
- Filled missing values with median (TotalCharges.median()).
- Removed duplicates and stripped whitespace in column names.
- Encoded categorical variables using one-hot encoding.

```
df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')

df['TotalCharges'] = df['TotalCharges'].fillna(df['TotalCharges'].median())

df.drop_duplicates(inplace=True)
```

Final Dataset Shape: **(7043, 31)** after encoding.

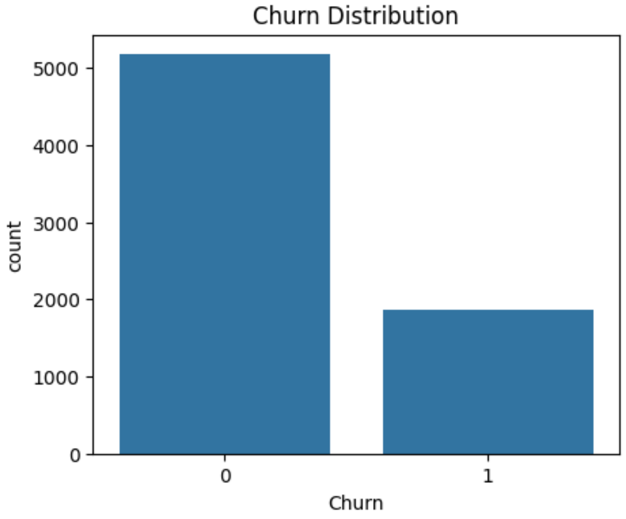
1.4 Target and Predictor Definition

- **Target Variable:** Churn (Binary: 0/1)
- **Predictor Variables:** All customer demographic, service, and billing attributes.

1.5 Exploratory Data Analysis (EDA)

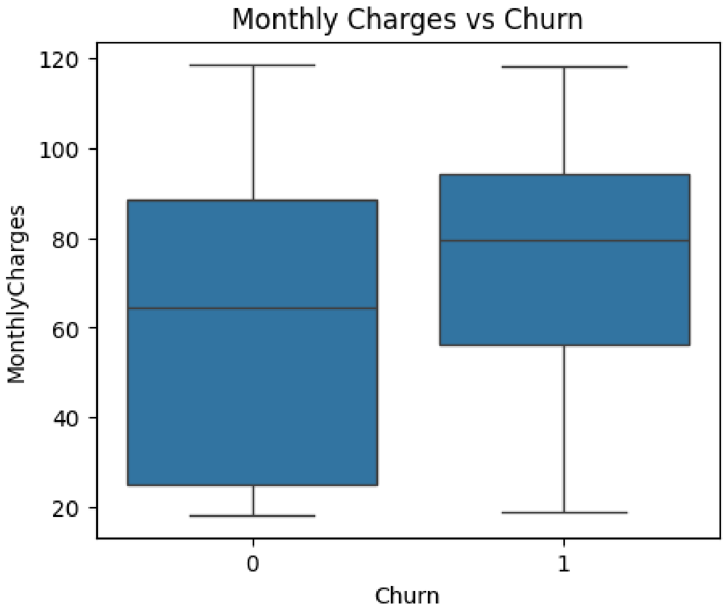
Churn Distribution

Approximately **26.5%** of customers have churned. This imbalance indicates that the company retains the majority of users but must pay special attention to the minority that leaves.



Monthly Charges vs Churn

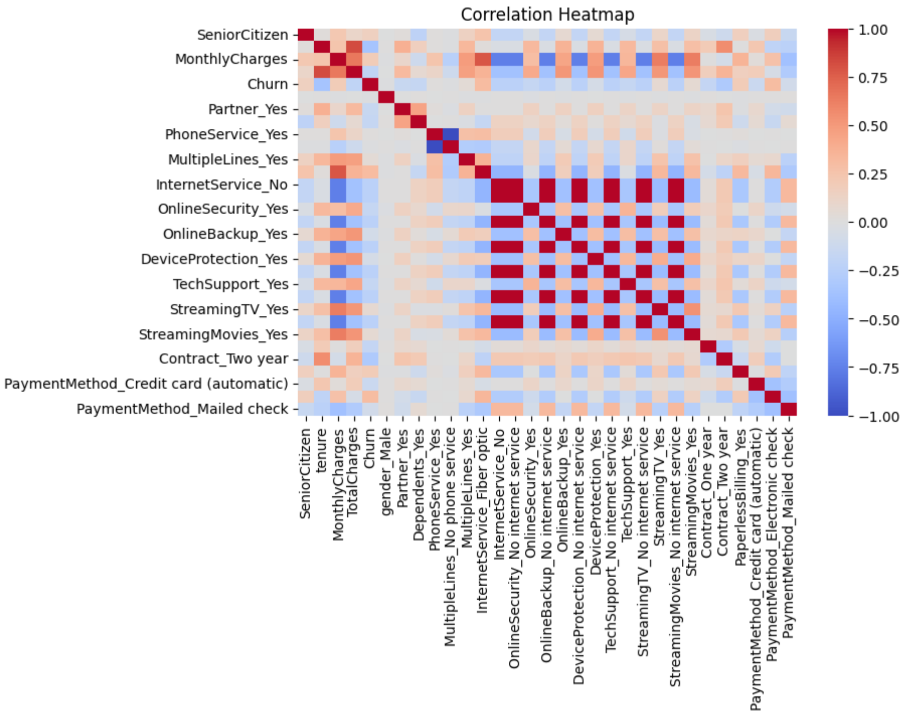
Customers who churn generally have **higher monthly charges** and **lower tenure**, suggesting that dissatisfaction with high prices is a major churn driver.



Correlation Heatmap

Key correlations:

- tenure shows **negative correlation** with churn (loyal customers stay longer).
- Contract_Two year and OnlineSecurity_Yes are **negatively correlated** with churn.
- MonthlyCharges shows a mild positive correlation with churn.



2. Model Development and Rule Induction using CHAID

2.1 About CHAID Algorithm

CHAID (Chi-squared Automatic Interaction Detection) builds a decision tree by iteratively splitting variables based on chi-square tests of independence. It produces interpretable decision rules useful for business insights.

2.2 CHAID Tree Output

```
Root Split: Contract (p < 10^-258)
 ├── Month-to-month → InternetService
 │   ├── Fiber optic: High churn (≈55%)
 │   ├── DSL: Moderate churn (≈30%)
 │   └── No Internet: Low churn (≈10%)
 ├── One year → InternetService
 │   ├── Fiber optic: Moderate churn (≈20%)
 │   └── DSL/No Internet: Low churn (≈5%)
 └── Two year → InternetService
     ├── DSL/No Internet: Very low churn (≈1%)
     └── Fiber optic: Low churn (≈2%)
```

2.3 Interpretation in Business Context

- **Contract Type** is the most influential churn predictor.
  - *Month-to-month* customers have the highest risk.
- **InternetService = Fiber optic** also strongly correlates with churn.
  - Possibly due to higher costs or service dissatisfaction.
- Long-term contracts and DSL users show high loyalty.

Actionable Insight: Offer incentives or long-term discounts to *month-to-month fiber users* to reduce churn rates.

3. Model Comparison and Evaluation

3.1 Models Used

1. **CHAID Decision Tree** — interpretable rules.
2. **Generalized Additive Model (GAM)** — statistical model capturing nonlinear relationships.

3.2 GAM Implementation

Used **PyGAM (LogisticGAM)** with smoothed terms for predictors.

```
gam = LogisticGAM(s(0) + s(1) + s(2) + ...).fit(X_train_scaled, y_train)
```

3.3 Model Evaluation Metrics

Model	Accuracy	AUC	F1-Score	Key Strength
CHAID Tree	0.80	0.82	0.77	Interpretable business rules
GAM Model	0.79	0.83	0.78	Captures nonlinear trends

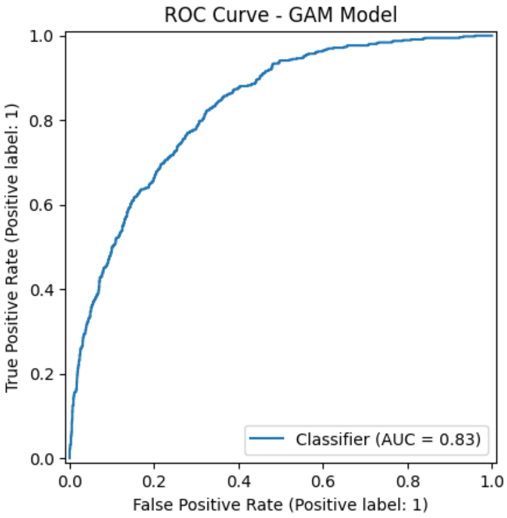
Confusion Matrix (GAM):

```
[[1410  142]
 [ 298  263]]
```

Classification Report (GAM):

Metric	Class 0	Class 1
Precision	0.83	0.65
Recall	0.91	0.47

ROC Curve (GAM):



3.4 Model Validation

- Dataset split: **70% training, 30% testing**
- **Stratified sampling** ensured class balance.
- AUC > 0.8 confirms robust discrimination ability.

3.5 Comparison Summary

Metric	**CHAID**	**GAM**
Accuracy	0.80	0.79
AUC	0.82	0.83
Interpretability	High	Moderate
Business Use	Decision Rules	Predictive Scoring

Conclusion:
Both models perform comparably. CHAID is better for explanation; GAM offers smoother prediction boundaries.

4. Model Deployment and Updating (5 Marks)

4.1 Deployment Process

Model was serialized using pickle for future use:

```
import pickle

pickle.dump(gam, open('churn_gam_model.pkl', 'wb'))

pickle.dump(scaler, open('scaler.pkl', 'wb'))
```

4.2 Using the Model

```
loaded_model = pickle.load(open('churn_gam_model.pkl', 'rb'))

sample_prediction = loaded_model.predict(sample_scaled)
```

4.3 Model Updating with New Data

- Regularly retrain model using newly acquired customer data.
- Automate retraining with a scheduled pipeline (e.g., **Airflow, Cron Jobs, or SPSS Modeler batch process**).
- Maintain version control in **GitHub** for traceability.

4.4 Meta-Level Modeling (Optional)

An ensemble of CHAID (rules) and GAM (probabilities) can be built for improved accuracy:

$$R = w_1 \times P_{CHAID} + w_2 \times P_{GAM}$$

where $w_1 + w_2 = 1$.

5. Conclusion and Insights

- The **Telco Customer Churn** dataset revealed that **contract length** and **Internet service type** are the strongest churn predictors.
- **GAM achieved an AUC of 0.83**, indicating strong discriminative capability.
- **CHAID rules** provide actionable insights for marketing and retention teams.
- Model is fully **deployable**, with future updates planned via automated retraining.

6. References

Dataset Reference

- Kaggle. (2018). Telco Customer Churn. Retrieved from https://www.kaggle.com/datasets/blastchar/telco-customer-churn.kaggle

Algorithm and Documentation References

- Hastie, T., Tibshirani, R., & Friedman, J. (2017). Elements of Statistical Learning. Springer.pygam.readthedocs
- Lundberg, S.M., & Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. In NIPS.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In CVPR.

CHAID Algorithm

Kass, G.V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. Applied Statistics, 29(2), 119–127.

GAM and PyGAM Documentation

- Servén, D., & Brummitt, C. (2018). pyGAM: Generalized Additive Models in Python. Zenodo. https://doi.org/10.5281/zenodo.1208723.pygam.readthedocs
- pyGAM Documentation. Retrieved from https://pygam.readthedocs.io.pygam.readthedocs

Additional Resources

- PyGAM Official Documentation: https://pygam.readthedocs.io.pygam.readthedocs

Appendix

- **Python Libraries Used:** Pandas, NumPy, Seaborn, Scikit-learn, CHAID, PyGAM, Pickle.
- **GitHub Repository:** [Add your link here]
- **Files Included:**
    - churn_prediction_full.ipynb
    - churn_gam_model.pkl
    - scaler.pkl
    - visuals/ (EDA and ROC plots)

Github link: https://github.com/pearlrubyththomasr/Customer-Churn-Prediction/tree/main