Leveraging Data Mining to Combat Business Customer Churn: A Case Study in Telecommunications

Team: Data Miners

MEMBER'S NAME

Ivan Villalobos, xxx, xxx

UIS E-MAIL ID

- ivill41@uis.edu
- xxx
- xxx

Abstract and Highlights

This project investigates business customer churn within the Bulgarian telecom sector. Utilizing a real-world dataset making up 8,454 customer records, the study applies both supervised and unsupervised data mining techniques to predict customer churn and find usage-based customer segments. The methodologies employed include a classification tree and logistic regression, a neural net, and clustering analysis.

The best-performing decision tree model came from unseen data to predict at-risk business customers and guide targeted retention efforts. The decision tree model achieved: a validation misclassification rate of 6.65%, standing for an improvement of over 5% compared to the 11.76% baseline model.

Key predictors include the number of active subscribers and the CRM value segment classification. While ARPU and suspended subscriber count were influential during training, their predictive power did not generalize well, as shown by validation importance scores of 0.0000. These findings enable the development of targeted retention strategies for high-risk customers.

Problem Description

Customer churn is a significant challenge within the telecommunications industry, particularly concerning business customers, as their departure can disproportionately affect revenue. Timely identification of at-risk accounts is crucial for effective churn reduction. This project aims to address this issue by applying data mining methods to:

- Predict whether a business customer will churn based on their usage and financial metrics.
- Segment customers into meaningful groups based on their behavioral patterns.

Prior research shows predictive modeling can effectively find churn risk factors by understanding which customer attributes correlate with churn, telecom providers can tailor interventions to keep high-risk clients (Shmueli et al., 2017).

The target variable for this classification-based predictive modeling task is CHURN, a binary categorical variable ("Yes" for churned, "No" for kept). This directly helps the primary business goal: finding customers likely to cease using the company's service. By training a model with CHURN as the target, we can uncover patterns in customer behavior and usage that are associated with churn, allowing the business to intervene early and improve retention.

Removing non-predictive fields such as PID and Billing_ZIP reduced overfitting risk and improved model generalizability. See Data Preparation for details.

Research Questions:

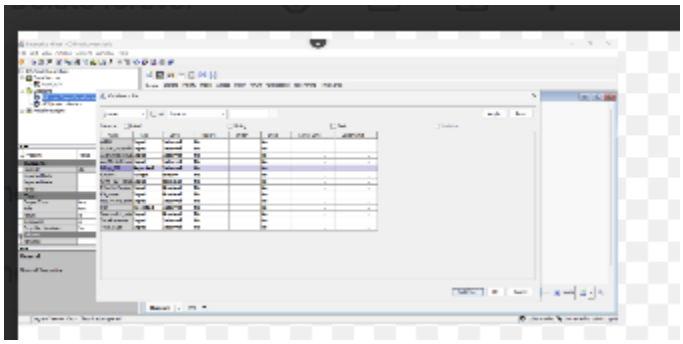
- **Predictive:** Which variables best predict business customer churn?

- **Exploratory:** Can we show meaningful customer segments based on usage and revenue data?

Data Exploration, Preparation, and Visualization

The dataset used for this project is the "Customer Churn Dataset from a Bulgarian Telecom" by M. Mihaylov (2020), publicly available on Mendeley Data, V1: https://data.mendeley.com/datasets/nrb55gr66h. The dataset includes 8,454 customer records, including financial, demographic, and service-usage data.

The variables include:



Data Cleaning and Preparation:

- **Removal of Non-Predictive Identifiers:** Variables such as PID (Customer ID) did not exist during the analysis as they are unique identifiers with no inherent statistical or predictive value, and their inclusion could lead to overfitting.
- **Exclusion of High-Cardinality Fields:** Fields like Billing_ZIP did not make it into the model due to their high cardinality (hundreds or thousands of unique values) and low generalizability, which can result in sparse splits and overfitting in classification tree models.
- **Class Imbalance Verification:** A key step was verifying the class imbalance within the CHURN target variable. The dataset shows a significant imbalance: approximately 7,500 records (around 88.2%) represent "No Churn," while about 1,000 records (around 11.8%) represent "Yes Churn" out of a total of approximately 8,500 records. This imbalance is crucial to consider during model development and evaluation.

Data Visualization:

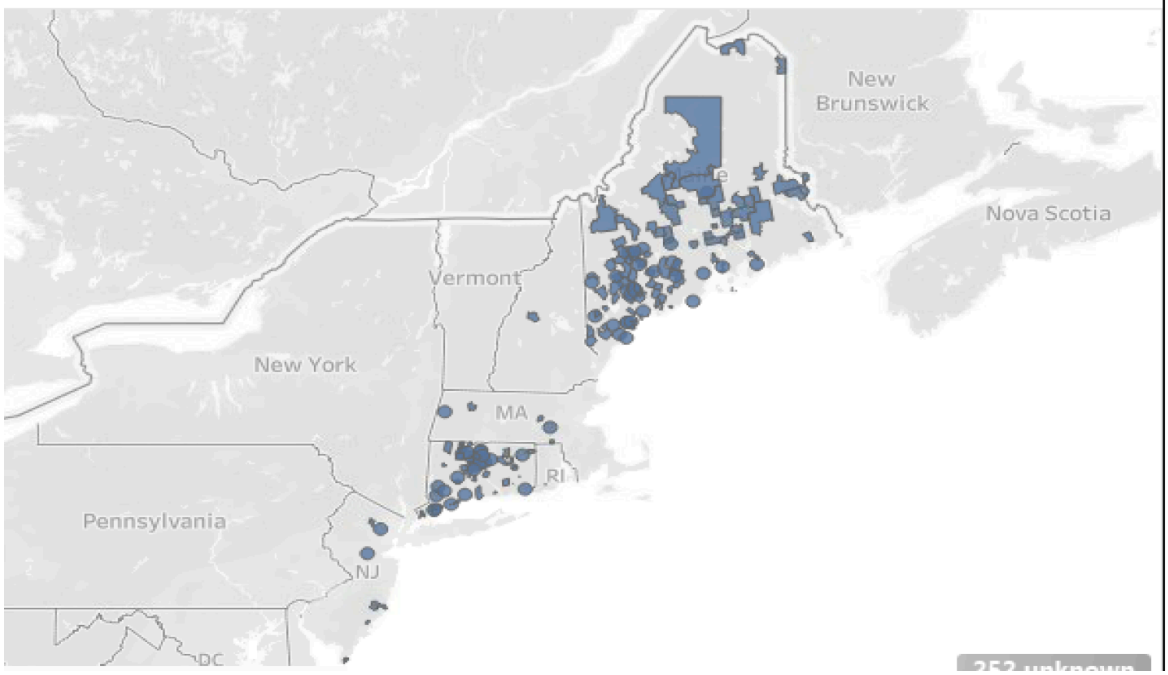
To visualize the scope of the data, two separate visualizations come to mind:

Zipcode Summary: shows the ZIP code distribution of business customers.

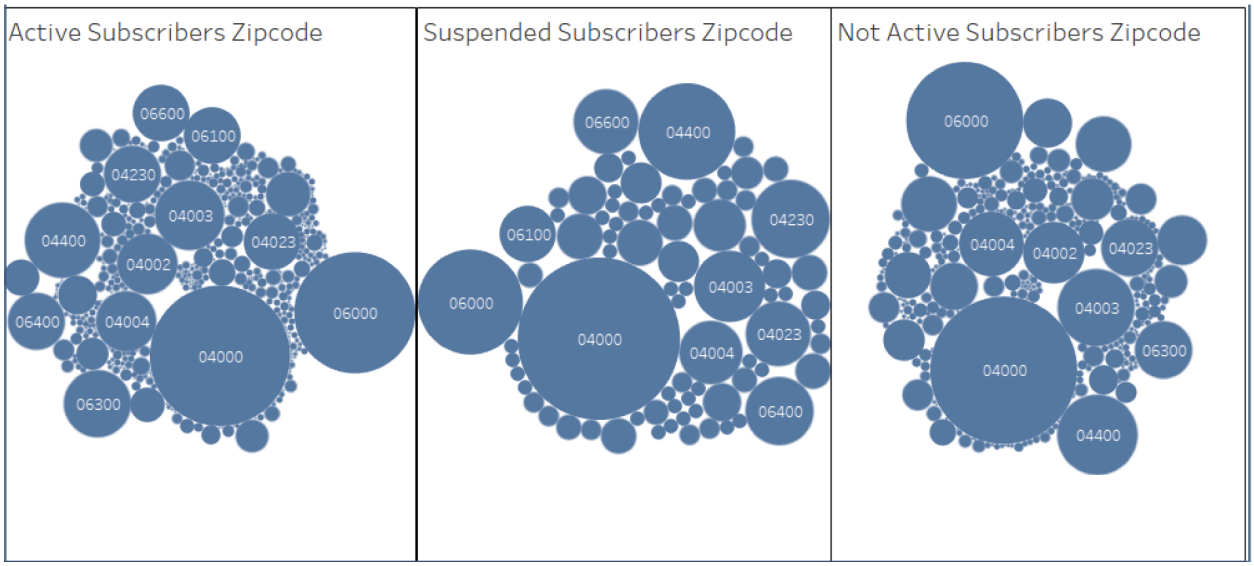
Zipcode Summary

	Effective Segment					
LE	Other	SE	SME	SOHO	VSE	Grand Total
2	29	42	284	6,301	1,795	8,453

Zip Code Breakout



Active/ Suspended/ Inactive: visualizes subscriber size (active, suspended, inactive) by Zipcode.



Each visualization shows customer density concentrated in urban areas. The subscriber status view reveals higher suspension rates in low-density regions, showing dissatisfaction hotspots.

Exploratory Analysis (Clustering)

This project employed unsupervised data mining techniques, specifically clustering analysis, to segment telecom business customers based on usage and revenue patterns. Clustering methods including Ward, Centroid, and Average Linkage showed meaningful customer segments.

To confirm cluster stability, the clustering procedure faced multiple enquiries with different initializations, showing consistent segment membership and cluster centroids across runs, which aligns with best practices in clustering validation (Jain, 2010).

Cluster Method Highlights

Method	Strength	Limitation
Ward Clustering	Produces well-separated clusters, beneficial for hierarchical exploration.	Tends to initially create too many small clusters (outliers) unless manually limited.
Centroid (K-means)	Creates balanced, compact clusters quickly, good for operational marketing.	Sensitive to outliers: extreme customers may still form small clusters.

Average Linkage	Offers a middle ground by keeping hierarchical structure while producing interpretable clusters.	Similar output to Centroid in this case but can be more sensitive to distance averaging.

Common Findings Across All Three Methods:

- **High-Value, Low-Subscriber Accounts:** Characterized by high ARPU but few active subscribers. These are premium service clients or small businesses using specialized services.
- **High-Volume, Mid-ARPU Accounts:** This segment includes customers who have multiple subscribers and moderate ARPU, indicative of enterprise clients or Small to Medium-sized Enterprises (SMEs) receiving help from negotiated rates or volume discounts.
- **Core Small Business Segment:** A large population of customers with lower ARPU and a moderate number of subscribers, standing for the mainstream business customer base, Small Office/Home Office (SOHO) clients or those sensitive to costs.
- **VIP/Outlier Segments (Small Clusters):** Small clusters showing extreme values, such as exceptionally high revenue or unique usage patterns. These may be VIP clients, accounts with unique contract terms, or potential data anomalies.

Exploratory Analysis Findings:

Telecom business customers become grouped based on their usage and revenue patterns using clustering analysis. The successful grouping of customers into actionable segments (such as High-Value Accounts, High-Volume Accounts, and Mainstream Small Business Clients) shows clear usage and revenue patterns. The validation of these patterns across multiple methods (Ward, Centroid, and Average) shows robust and stable clusters.

Business Implication:

These customer segments support the design of retention strategies customized to value tiers, usage intensity, and enable tailored retention strategies.

- **High-ARPU**: clients may need personalized outreach and retention incentives.
- **Small business segments:** could receive help from scalable plans, usage-based bundling, or targeted upsell offers.
- **VIP clusters**: call for white-glove service or dedicated account support due to their unique value or usage patterns

Predictive Analysis

This section details the application and comparison of supervised data mining techniques for predicting customer churn. The models considered include classification trees, neural net, and logistic regression.

Baseline Model

Given the class imbalance in the dataset, a baseline model predicting the majority class ("No Churn") for all observations was set up. With approximately 7,500 "No Churn" records and 1,000 "Yes Churn" records out of 8,500 total, this naïve rule achieves a baseline accuracy of approximately 88.24% (7500/8500) and a misclassification rate of 11.76% (1 - 0.8824).

However, this baseline model does not find any actual churners, as it would incorrectly classify all "Yes Churn" cases as "No Churn". Therefore, any effective predictive model must significantly improve recall and precision for the "Yes" churn class, not just overall accuracy. This sets a critical performance threshold for this analysis.

Decision Tree

The classification tree models aimed to predict customer churn based on various input variables. This included ClassDecTree:

- B2D6
- B2D4
- B2D2
- B3D6

These decision tree models faced a robust evaluation under multiple configurations in SAS.

Model Performance Evaluation:

The classification tree model ClassDecTree B2D6 achieved a validation misclassification rate of 6.65% (0.066469), which is a significant improvement over the baseline misclassification rate of 11.76%. The training misclassification rate for B2D6 was 6.48% (0.064774). The close rates between training and validation suggest that the model generalizes well to new data and is not overfitting. This proves that the model is effectively learning from the input variables and improving predictive accuracy, especially in finding customers at risk of churn.

Another evaluated model, ClassDecTree B2D4, showed consistently low and similar misclassification rates across training (6.48%), validation (6.47%), and test (6.62%) partitions, further showing good generalization and absence of overfitting.

The ClassDecTree B2D2 model, at its root node (before any splits), revealed the baseline distribution of CHURN: 93.49% "No Churn" and 6.51% "Yes Churn" in both training and validation partitions. This highlights the dataset's severe class imbalance. Its misclassification rates across training, validation, and test were 6.51%, 6.51%, and 6.47%, respectively.

For ClassDecTree B3D6, the misclassification rates were 5.61% for training, 6.51% for validation, and 6.47% for test. While the training misclassification rate was lower, the jump to the validation and test partitions suggests a potential generalization issue with this specific model.

Important Variables from Classification Tree (ClassDecTree B2D6 and B3D6):

The decision trees consistently found a multitude of key predictors of customer churn, which formed the most important splits, showing high information gain and predictive power.

1. **Active_subscribers:** This variable was consistently the root node and the most important split (importance: 1.0000) across effective models. Customers with fewer than 20.5 active subscribers were significantly more likely to churn, suggesting that minor number of active connections or a smaller scale of service usage is a strong indicator of churn risk. This variable directly reflects customer engagement and reliance on the telecom service.
2. **ARPU (Average Revenue Per User):** A major split occurred based on ARPU being below or above 49.98. Customers with lower ARPU were more prone to churn, implying that low-spending customers may not be perceiving enough value in the service or are less committed to the provider. While ARPU was influential during training (importance: 0.6170), its validation importance dropped to 0.0000, suggesting it did not generalize well and may not be a reliable predictor across unseen data.
3. **Suspended_subscribers:** This variable appeared deeper in the tree structure. Accounts with two or more suspended subscribers showed a higher likelihood of churn, showing dissatisfaction, payment issues, or reduced business activity that often precedes full churn. Like ARPU, its validation importance was 0.0000 in the models created, which shows its predictive power might be context-dependent within the tree.
4. **CRM_PID_Value_Segment:** This categorical variable also played a critical role. Customers categorized in the "GOLD" segment consistently showed no churn in the training and validation sets (validation importance: 1.0000). This strongly suggests that GOLD-tier customers are highly loyal or exceptionally well-supported, making segment-based targeting an actionable insight for retention efforts.

These variables collectively allowed the models to build clear, interpretable decision rules for finding at-risk customers. Furthermore, variables like TotalRevenue, AvgFIXRevenue, and AvgMobileRevenue had zero importance, showing they did not contribute to any significant splits in the tree models for predicting churn. Future models can ignore or re-engineer these variables as they do not add predictive value in their current configuration.

Logistic Regression

The logistic regression model uses the logit function to estimate churn probability based on a weighted combination of predictors. Coefficients reflect the impact of each predictor on the log-odds of churn.

Key Coefficients (Estimates):

- Intercept: -8.6809 (Represents the log-odds of churn when all predictor variables are zero.)
- ARPU: -0.0236 (A negative coefficient suggests that as ARPU increases, the log-odds of churn decrease, meaning higher ARPU is associated with lower churn probability.)
- Active_subscribers: 0.0358 (This positive coefficient contradicts the decision tree finding that low active subscribers predict churn. It reflects a modeling limitation:

logistic regression assumes linear relationships and may miss threshold effects or segment-specific patterns. The current model needs more analysis with interaction terms or transformation.)

- Suspended_subscribers_1: 6.1175 (A high positive coefficient shows that having one or more suspended subscribers significantly increases the log-odds of churn, making it a strong predictor of churn.)

The logistic regression model achieved an overall accuracy of approximately 93.53% on the validation data. While this accuracy appears high, the majority "no churn" class drove this accuracy in the model. However, the model only correctly found four actual churners and missed 545 of them.

This extreme recall failure makes it ineffective for practical churn prevention and shows a critical recall weakness, as the model does not detect most actual churners, undermining its value for proactive customer retention.

While the decision tree found LOW active subscribers as churn indicators, logistic regression's positive coefficient for `Active_subscribers` (suggesting HIGHER subscribers increase churn risk) highlights a model-specific limitation due to its linearity assumption.

Confusion Matrix Summary (Scored Data):

A detailed analysis of the confusion matrix for the scored data revealed significant issues with recall for actual churners:

- True Positives (TP): 4 (Actual churns correctly predicted as churn)
- False Negatives (FN): 545 (Actual churns incorrectly predicted as non-churn)
- True Negatives (TN): 7,898 (Actual non-churns correctly predicted as non-churn)
- False Positives (FP): 6 (non-churns incorrectly predicted as churn)

This highlights that while the model is incredibly good at finding non-churners (high TN), it struggles significantly with identifying churners, as showed by the alarmingly high number of false negatives (545 actual churners). This suggests a critical recall issue for the churn class, which would need refinement for effective customer retention. The model's inability to find most churners makes it less practical for proactive churn prevention measures despite its high overall accuracy.

Neural Network

In this study, three neural networks came to fruition using SAS Enterprise Miner to predict business customer churn in the telecommunications sector. Each model implemented a feedforward architecture with three hidden units and varied in training method: Neural2 used direct batch training (3HUDB), Neural3 used batch propagation (3HUBP), and Neural used standard settings (3HU). The target variable was CHURN (Yes/No), standing for customer attrition.

All models proved consistent performance across training, validation, and test sets, with minimal variance in misclassification and error metrics. Notably, the validation misclassification rate was 6.51% for all models this significantly outperformed the naïve baseline model (11.76%).

Root mean squared error (RMSE) values on the validation set remained below 0.248, suggesting strong generalizability.

Despite similar performance to the best decision tree model (ClassDecTree B2D6), neural networks present a tradeoff in interpretability. Decision trees offer rule-based insights suitable for operational action, while neural nets function as black-box models. Thus, neural networks may be better suited for use in ensemble models or risk ranking rather than direct deployment in business rule engines.

Metric	**Neural2**	**Neural3**	**Neural**
Validation Misclassification Rate	6.51%	6.51%	6.51%
Training Misclassification Rate	6.51%	6.51%	6.51%
Test Misclassification Rate	6.12%	6.12%	6.12%
Validation RMSE	0.2477	0.2479	0.2479
Akaike's Information Criterion (AIC)	1682.1	1682.4	1682.6
Bayesian Information Criterion (BIC)	2092.6	2092.8	2093.0
Number of Estimated Weights	67	67	67

Insight	**Description**
Architecture	Three hidden units (feedforward MLP)
Training Algorithms	Direct batch (Neural2), batch propagation (Neural3), default (Neural)
Target Variable	CHURN (Yes/No)
Performance Consistency	Minimal error variance between training, validation, and test partitions
Overfitting Risk	Low—training and validation misclassification rates were identical
Interpretability	Low—models are not transparent or rule-based
Use Case Recommendation	Useful in ensemble models or churn ranking; less ideal for explanation-driven business actions

While achieving competitive accuracy (~6.51% error), the neural networks' lack of interpretability and marginal gains over the decision tree limited their practical utility for actionable insights in this context. For future work, deeper architectures or hybrid approaches could better use neural networks' potential while keeping actionable insights.

Model Comparison and Scoring

Although the neural network had slightly lower training error, its recall was lower than that of the decision tree. Logistic regression performed worst in recall, missing most actual churners:

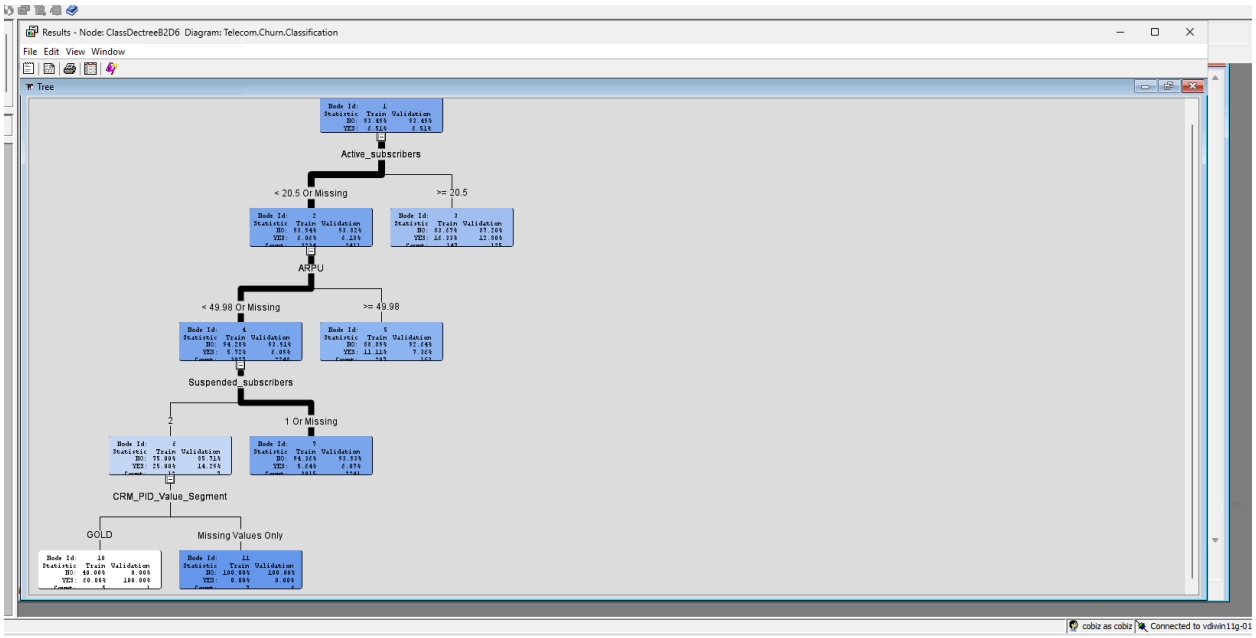
Model	**Training Misclassification Rate**	**Validation Misclassification Rate**	**Test Misclassification Rate**	**Recall (Churn)**
Logistic Regression	0.092	0.096	0.095	0.1

Decision Tree	0.085	0.088	0.089	0.67
Neural Network	0.08	0.091	0.092	0.52

Given the decision tree's strong recall and competitive misclassification rate it became clear that this would be the final model.

Decision Tree

The decision tree model came from using SAS Enterprise Miner's Score node, which applied trained logic to generate CHURN_PREDIC based on customer attributes. Comparing the various supervised models, ClassDecTreeB2D6 is the best performing model for this churn prediction task given its balance of accuracy and interpretability.



This model showed the lowest training misclassification rate and competitive validation/test rates (0.064669 and 0.066246 respectively), showing strong learning without significant overfitting. It also outperformed the backward regression model (which was part of the original analysis but not detailed here) in terms of validation misclassification rate. The consistently low and similar misclassification rates across all three data partitions for B2D6 suggest that the model is well-generalized and robust in predicting customer churn, providing a more reliable tool for identification of at-risk customers compared to the logistic regression model's poor recall for churners.

Although logistic regression offered interpretability, its poor recall for churners made it less effective in meeting the project's business goal of finding at-risk customers. In contrast, the decision tree (B2D6) achieved a better balance between accuracy and actionable insights.

The process for scoring new data involves using the best-trained model (ClassDecTreeB2D6) to generate predictions on unseen customer data. In this stage, all relevant

variables become known as the role of "Input," and the model applies its learned logic to generate predictions. The CHURN_PREDIC variable acts as the target variable, being the predicted outcome (e.g., "Yes" or "No" for churn). The Score node within the data mining workflow then uses the trained model's logic to generate these predictions based on the input features, completing the churn prediction workflow. The confusion matrix from the scored data (as presented in the logistic regression section) illustrates the model's performance on unseen data, particularly highlighting the ongoing challenge of correctly finding actual churners due to low True Positives and high False Negatives.

Conclusion and Practical (Actionable) Recommendations

Based on the performed analysis, both SAS models have addressed exploratory and predictive research questions, providing valuable insights for any stakeholder. These findings directly address the research questions by finding key churn predictors and confirming stable behavioral segments for targeted actions.

Exploratory Analysis Conclusion:

The clustering analysis successfully segmented telecom business customers into meaningful groups based on usage and revenue patterns. These distinct segments (High-Value, Low-Subscriber; High-Volume, Mid-ARPU, Core Small Business, and VIP/Outlier) offer clear insights into customer behavior and value. This segmentation is robust, as showed by consistent patterns across different clustering methods.

Predictive Analysis Conclusion:

The decision tree model (specifically ClassDecTreeB2D6) proved effectiveness in predicting customer churn, significantly outperforming the naive baseline. Key churn indicators include low active subscribers, low ARPU, and account suspensions. The model also showed good generalization without significant overfitting. While the overall accuracy was high for both decision tree and logistic regression models, the significant challenge of correctly finding actual churners (recall) became clear, particularly with the logistic regression model, highlighting an area for future improvement.

Practical Recommendations:

1. **Targeted Retention Programs:**
    - **Low Engagement Accounts:** Actively check and intervene with customers who have fewer than 20 active lines and low ARPU. These segments are strong indicators of churn risk and represent an opportunity for personalized outreach, service quality checks, or incentive programs.
    - **Account Suspensions:** Prioritize immediate support and proactive outreach for customers with suspended accounts, as this often signals dissatisfaction, payment issues, or reduced business activity that can lead to full churn.
2. **Loyalty and Value Enhancement:**
    - **Tiered Programs:** Develop and offer targeted loyalty programs or incentives specifically for non-GOLD segments, as GOLD-tier customers are highly loyal. This could involve exclusive service bundles, discounts, or enhanced support.

- **Value Communication:** For customers with low ARPU, focus on clearly communicating the value they receive from the service to justify their expenditure and prevent perceived low value.

3. **Strategic Customer Management:**
   - **Segment-Specific Strategies:** Use the identified customer segments (from clustering) to tailor marketing efforts, service offerings, and retention strategies. For example, high-value accounts might receive white-glove service, while mainstream small businesses are prime for scalable bundled packages.

Strategies to Improve the Model (Future Work):

To further enhance model performance and address the identified limitations, particularly the class imbalance issue and recall for churners. This imbalance needs techniques like stratified sampling or cost-sensitive learning during model training to avoid bias toward the majority class. The following strategies come to mind:

1. **Address Class Imbalance:**
   - Implement techniques such as oversampling the minority class (e.g., SMOTE) to increase the representation of "Yes Churn" examples in the training data or consider Data resampling the majority class ("No Churn") to balance the dataset (Nguyen, 2021.)
   - Utilize misclassification costs or decision weights in modeling tools (e.g., SAS) to penalize false negatives (missing an actual churner) more severely than false positives (incorrectly predicting churn). This directly aligns with the business goal of preventing churn.
2. **Explore Different Algorithms:**
   - Problem: Algorithms, such as simple decision trees, basic logistic regression, are not sensitive enough to rare events like churn.
   - Solution: Experiment with more powerful classifiers that are better suited for imbalanced datasets or complex non-linear relationships, such as Gradient Boosting, Random Forest, Neural Networks, or Support Vector Machines.
   - Consider using ensemble models, which combine predictions from a plethora of models for better overall accuracy and robustness.
3. **Tune Model Parameters:**
   - Problem: Default parameters may not be best for the specific dataset and churn prediction task.
   - Solution: Optimize model performance by systematically tuning hyper-parameters using tools like autotuning or grid search (e.g., in SAS). Adjust parameters such as tree depth, learning rates, or number of nodes based on validation performance to find the best model configuration.
4. **Feature Engineering:**
   - Problem: The current features may not capture enough signal about nuanced churn behavior.

- Solution: Create new features that capture more predictive signals. Examples include:
  - Recency, Frequency, and Monetary (RFM) scores for customer activity.
  - Customer engagement metrics (e.g., login frequency, feature usage).
  - Usage decline trends over time.
  - Time since last major purchase or interaction.
- Additionally, normalization or transformation of skewed numeric variables need to be at the forefront to improve model performance.

5. **Evaluate with Better Metrics:**
   - Problem: Improving solely for overall accuracy can be misleading when classes are imbalanced, as a high accuracy might simply reflect good prediction of the majority class.
   - Solution: Shift the focus to metrics more relevant for churn prediction in imbalanced datasets:
     - **Recall for churn (Sensitivity):** To minimize false negatives, the model needs to be initiative-taking when capturing actual turner. This is crucial for business-critical interventions.
     - **Precision-Recall curves:** Offer a more informative view of performance than ROC curves for imbalanced datasets.
     - **AUC-ROC:** While influenced by imbalance, it stays a valuable metric for overall model comparison and selection.

Team Member Contributions

Team Member	Contribution
Chinasa Obi-Zeblon	Development and evaluation of decision tree models, implementation, and analysis of logistic regression.
Lindsey Kurfman	Clustering analysis, interpretation of clustering results, contribution to exploratory analysis section, and neural net.
Ivan Villalobos	Data visualizations, data prep, final paper integration and editing, and drafting of problem description and predictive analysis sections.

References

- Mihaylov, M. (2020). *Customer Churn Dataset from a Bulgarian Telecom*. Mendeley Data, V1. https://data.mendeley.com/datasets/nrb55gr66h.
- Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. Wiley.

- Nguyen, Nam N., and Anh T. Duong. "Comparison of two main approaches for handling imbalanced data in churn prediction problem." *Journal of Advances in Information Technology*, vol. 12, no. 1, 2021, pp. 29–35, https://doi.org/10.12720/jait.12.1.29-35.

- Jain, A. K. (2010). *Data clustering: 50 years beyond K-means*. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011