Babu Banarasi Das University

Predictive Analysis
(BCADS15301)

A Comprehensive Case Study
on
Data Mining and Integration of Telecommunication Dataset for Predicting Customer Churn Using IBM SPSS Modeler

SUBMITTED TO: Mr. VIKESH SIR

SUBMITTED BY:
Name: NEHA
Roll No: 1230258275
Class: BCADS34

Date: November 06, 2025

ABSTRACT

Customer churn represents a critical challenge in the telecommunications industry, directly impacting revenue and market share. This project presents a comprehensive data mining case study focused on integrating and analyzing telecommunication datasets to predict customer churn patterns using IBM SPSS Modeler. The study demonstrates systematic data preparation methodologies including data cleaning, merging multiple data sources, feature engineering, and implementation of predictive modeling techniques. Through the application of CHAID (Chi-square Automatic Interaction Detection) decision tree algorithms and strategic data manipulation operations, this research identifies key factors influencing customer attrition and develops a robust framework for identifying high-risk customers. The methodology emphasizes practical implementation using SPSS Modeler's visual programming interface, making advanced analytics accessible for business intelligence applications. Results demonstrate the effectiveness of integrated data mining approaches in transforming raw telecommunication data into actionable insights for customer retention strategies.

1. INTRODUCTION

1.1 Background and Context

In the contemporary telecommunications landscape, customer churn has emerged as one of the most pressing challenges facing service providers. With market saturation and intense competition, acquiring new customers costs significantly more than retaining existing ones. Understanding and predicting customer churn enables organizations to implement proactive retention strategies, optimize marketing investments, and maintain competitive advantage. This project addresses these challenges through a systematic data mining approach utilizing IBM SPSS Modeler, a powerful visual data science platform.

1.2 Problem Statement

Telecommunication companies maintain vast repositories of customer data across multiple systems, including demographic information, service usage patterns, and billing records. However, this data remains fragmented and underutilized for predictive analytics. The primary challenge is to integrate these disparate data sources, clean and prepare the data, and develop predictive models that can accurately identify customers at risk of churning. This project demonstrates a complete workflow from data integration to actionable customer risk assessment.

1.3 Objectives

The primary objectives of this case study are:

- To demonstrate effective data integration techniques for combining multiple telecommunicationdatasets
- To implement comprehensive data quality assurance and cleaning procedures
- To develop predictive models using CHAID decision tree algorithms for churn prediction
- To identify and segment high-risk customers requiring targeted retention interventions
- To establish a reproducible analytical workflow using IBM SPSS Modeler's visual programmingparadigm

1.4 Scope and Methodology

This project employs IBM SPSS Modeler as the primary analytical tool, leveraging its visual data mining capabilities. The methodology encompasses data importation, merge operations, data validation, feature engineering, predictive modeling using CHAID algorithms, and customer risk segmentation. The approach emphasizes practical implementation techniques applicable to real-world business intelligence scenarios.

2. METHODOLOGY AND IMPLEMENTATION

The implementation follows a systematic workflow utilizing IBM SPSS Modeler's node-based visual programming interface. Each stage represents a critical component of the data mining pipeline, from raw data acquisition through model deployment and customer risk assessment.

2.1 Dataset Description

The analysis utilizes two primary datasets:

Customer_Info.csv: Contains demographic and account information including customer identifiers, gender, senior citizen status, partner and dependent information, tenure duration, and service subscriptions.

Usage_Billing.csv: Encompasses billing and usage metrics including monthly charges, total charges, payment methods, contract types, and historical churn indicators.

These datasets are linked through a common CustomerID field, enabling comprehensive customer profiling by combining behavioral, demographic, and financial attributes.

2.2 Analytical Tool: IBM SPSS Modeler

IBM SPSS Modeler provides a comprehensive visual data science environment that enables users to build predictive models without extensive programming knowledge. The platform utilizes a drag-and-drop interface where analytical operations are represented as nodes connected in streams, creating transparent and reproducible analytical workflows. This approach democratizes advanced analytics, making sophisticated data mining techniques accessible to business analysts and domain experts.

2.3 Implementation Workflow

The following sections detail the complete implementation process, illustrating each analytical operation with corresponding visual representations from SPSS Modeler.

Step 1: Data Source Configuration and Import

The initial phase involves configuring data sources within SPSS Modeler. Two Excel source nodes are instantiated on the modeling canvas, each configured to import one of the primary datasets (Customer_Info.csv and Usage_Billing.csv). The source nodes are configured with appropriate file paths, data type specifications, and field definitions. A Table output node is connected to each source to verify successful data importation and perform initial data quality inspection.

Technical Implementation: *Navigation: Source Palette → Excel Node → Configure File Path → Connect to Table Node*

INDEX

Sr. No	Name of Experiment	Date	Faculty Signature	Remarks
1	DEMONSTRATE THE USE OF NODES IN SPSS FOR IMPORTING A DATASET, APPLYING DATA FILTERS TO SELECT SPECIFIC RECORDS, AND EXPORTING THE FILTERED RESULTS	05/09/25		
2	DEMONSTRATE HOW TO USE NODES IN SPSS TO GATHER INITIAL DATA FOR A TELECOMMUNICATIONS COMPANY.	10/09/25		
3	CREATE A DATA-MINING PROJECT TO PREDICT CHURN IN TELECOMMUNICATIONS FIRM.	11/09/25		
4	CREATE A DATA-MINING PROJECT TO PREDICT CHURN IN TELECOMMUNICATIONS FIRM	13/09/25		
5	DEFINE THE UNIT OF ANALYSIS FOR THE TELECOMMUNICATIONS DATASET IN SPSS USING NODES.	15/09/25		
6				
7				
8				

Figure 1: Step 1: Data Source Configuration and Import

Step 2: Initial Data Verification

Following data importation, preliminary data inspection is performed using Table output nodes. This verification step ensures that data has been correctly loaded, field names are properly recognized, data types are accurate, and the expected number of records has been imported. This quality checkpoint is essential before proceeding with data integration operations.

Technical Implementation: *Operation: Output → Table Node → Execute → Review Data Structure*

INDEX

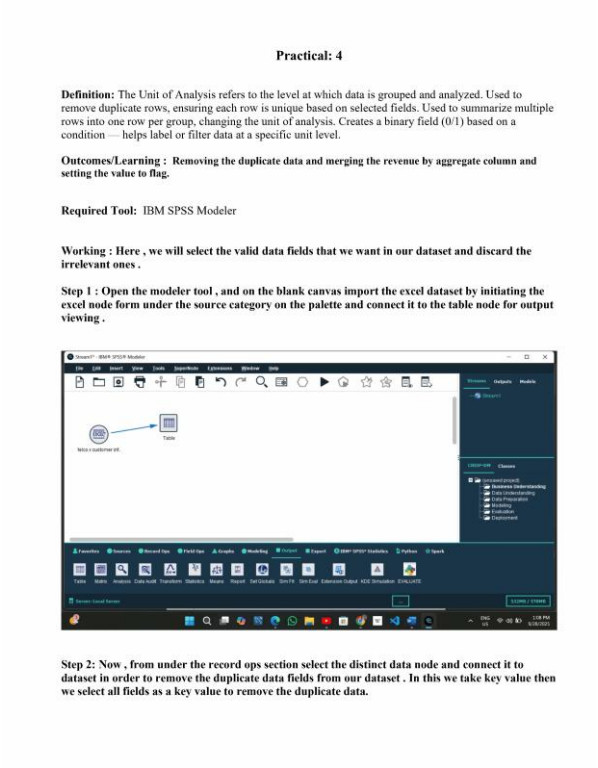
Sr. No	Name of Experiment	Date	Faculty Signature	Remarks
1	DEMONSTRATE THE USE OF NODES IN SPSS FOR IMPORTING A DATASET, APPLYING DATA FILTERS TO SELECT SPECIFIC RECORDS, AND EXPORTING THE FILTERED RESULTS	05/09/25		
2	DEMONSTRATE HOW TO USE NODES IN SPSS TO GATHER INITIAL DATA FOR A TELECOMMUNICATIONS COMPANY.	10/09/25		
3	CREATE A DATA-MINING PROJECT TO PREDICT CHURN IN TELECOMMUNICATIONS FIRM.	11/09/25		
4	CREATE A DATA-MINING PROJECT TO PREDICT CHURN IN TELECOMMUNICATIONS FIRM	13/09/25		
5	DEFINE THE UNIT OF ANALYSIS FOR THE TELECOMMUNICATIONS DATASET IN SPSS USING NODES.	15/09/25		
6				
7				
8				

Figure 2: Step 2: Initial Data Verification

Step 3: Data Merge Configuration

Data integration is accomplished through the Merge node from the Record Operations palette. This node combines the two datasets using CustomerID as the key field. The merge operation is configured as a partial outer join to include all matching records plus selected non-matching records, ensuring comprehensive data coverage while maintaining referential integrity. This approach preserves customers who may exist in one dataset but not the other, enabling complete analysis.

> ***Technical Implementation:*** *Configuration: Record Ops → Merge Node → Select Keys Method → Choose CustomerID → Partial Outer Join*



Practical: 4

Definition: The Unit of Analysis refers to the level at which data is grouped and analyzed. Used to remove duplicate rows, ensuring each row is unique based on selected fields. Used to summarize multiple rows into one row per group, changing the unit of analysis. Creates a binary field (0/1) based on a condition — helps label or filter data at a specific unit level.

Outcomes/Learning : Removing the duplicate data and merging the revenue by aggregate column and setting the value to flag.

Required Tool: IBM SPSS Modeler

Working : Here , we will select the valid data fields that we want in our dataset and discard the irrelevant ones .

Step 1 : Open the modeler tool , and on the blank canvas import the excel dataset by initiating the excel node form under the source category on the palette and connect it to the table node for output viewing .

Step 2: Now , from under the record ops section select the distinct data node and connect it to dataset in order to remove the duplicate data fields from our dataset . In this we take key value then we select all fields as a key value to remove the duplicate data.

Figure 3: Step 3: Data Merge Configuration

Step 4: Merge Operation Parameters

The Merge node configuration requires careful specification of merge parameters. From the available fields, CustomerID is selected as the merge key from both datasets. The merge type is set to "Include matching and selected non-matching records" (partial outer join), which retains all customers with complete information while preserving those with partial data. This configuration balances data completeness with analytical validity.

> ***Technical Implementation:*** *Parameters: Merge Method = Keys | Key Field = CustomerID | Type = Partial Outer Join*

Practical: 4

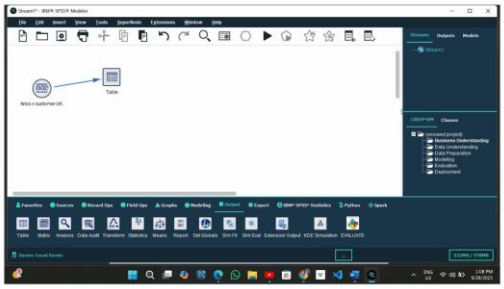
Definition: The Unit of Analysis refers to the level at which data is grouped and analyzed. Used to remove duplicate rows, ensuring each row is unique based on selected fields. Used to summarize multiple rows into one row per group, changing the unit of analysis. Creates a binary field (0/1) based on a condition — helps label or filter data at a specific unit level.

Outcomes/Learning : Removing the duplicate data and merging the revenue by aggregate column and setting the value to flag.

Required Tool: IBM SPSS Modeler

Working : Here , we will select the valid data fields that we want in our dataset and discard the irrelevant ones .

Step 1 : Open the modeler tool , and on the blank canvas import the excel dataset by initiating the excel node form under the source category on the palette and connect it to the table node for output viewing .



Step 2: Now , from under the record ops section select the distinct data node and connect it to dataset in order to remove the duplicate data fields from our dataset . In this we take key value then we select all fields as a key value to remove the duplicate data.

Figure 4: Step 4: Merge Operation Parameters

Step 5: Post-Merge Data Validation

After executing the merge operation, the integrated dataset is examined using a Table output node. This validation step confirms successful data integration, verifies that fields from both source datasets are present, checks for correct record counts, and identifies any data quality issues introduced during the merge process. This intermediate verification is crucial for ensuring data integrity throughout the analytical pipeline.

> ***Technical Implementation:*** *Validation: Connect Table Node to Merge Output → Execute → Verify Field Integration*

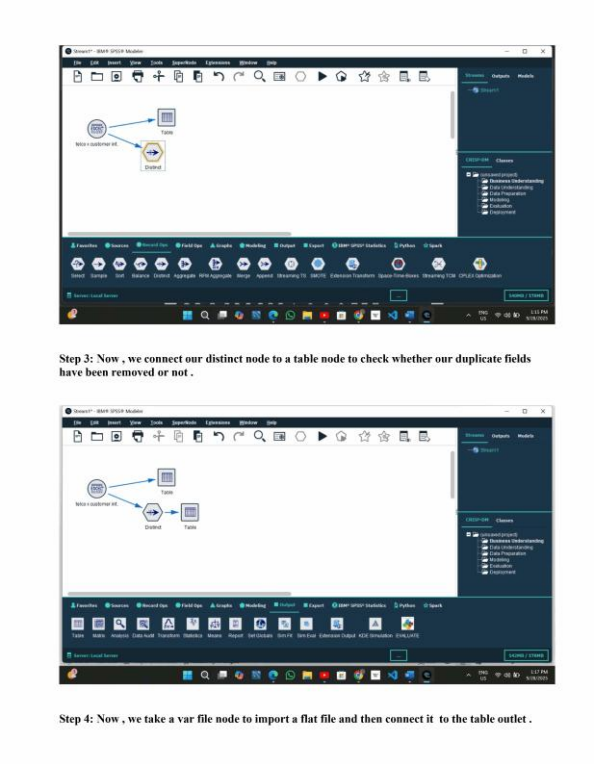
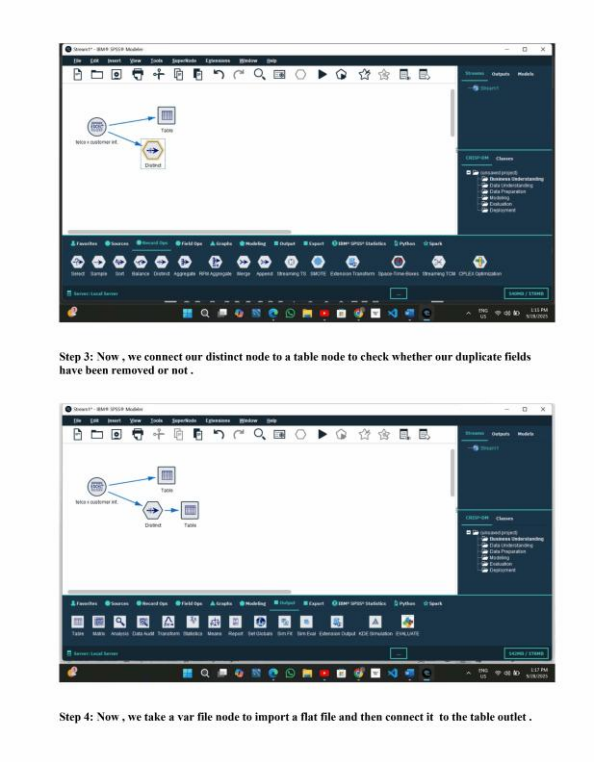


Figure 5: Step 5: Post-Merge Data Validation

Step 6: Data Filtering and Quality Assurance

A Select node is introduced to implement comprehensive data quality filters. This node applies logical conditions to exclude invalid records, ensuring that only complete and valid data proceeds to the modeling phase. The filtering criteria are designed to eliminate records with missing churn indicators, zero or negative monetary values, and insufficient tenure information, all of which would compromise model accuracy.

> ***Technical Implementation:*** *Node: Record Ops → Select Node → Connect to Merge Output*



Step 3: Now , we connect our distinct node to a table node to check whether our duplicate fields have been removed or not .

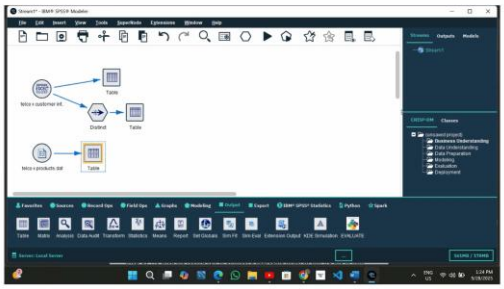
Step 4: Now , we take a var file node to import a flat file and then connect it to the table outlet .

Figure 6: Step 6: Data Filtering and Quality Assurance

Step 7: Filter Condition Specification

The Select node is configured with sophisticated filtering logic to ensure data quality. The filter expression implements multiple conditions: exclusion of null churn values [not(@NULL(Churn))], removal of empty churn indicators [not(Churn="")], validation of positive monthly charges [MonthlyCharges > 0], verification of positive total charges [TotalCharges > 0], and confirmation of positive tenure [Tenure > 0]. These conditions work in conjunction to create a clean, analysis-ready dataset.

> ***Technical Implementation:*** *Filter Logic: not(@NULL(Churn)) AND not(Churn="") AND MonthlyCharges>0 AND TotalCharges>0 AND Tenure>0*



Step 5: Under the record ops section , to connect a aggregate node we use to sum the revenue field values. Then we take key value as revenue and apply sum and mean .

Step 6: Now , we will connect the table node to see the output in tabular format .

Figure 7: Step 7: Filter Condition Specification

Step 8: Filtered Data Verification

Following the application of data quality filters, a Table node is connected to verify the filtering results. This verification confirms that invalid records have been successfully removed, assesses the impact of filtering on dataset size, and ensures that remaining records meet all quality criteria. This checkpoint validates that the data cleaning process has been effective and that the dataset is suitable for modeling.

> ***Technical Implementation:*** *Verification: Connect Table Node → Execute → Review Filtered Records*

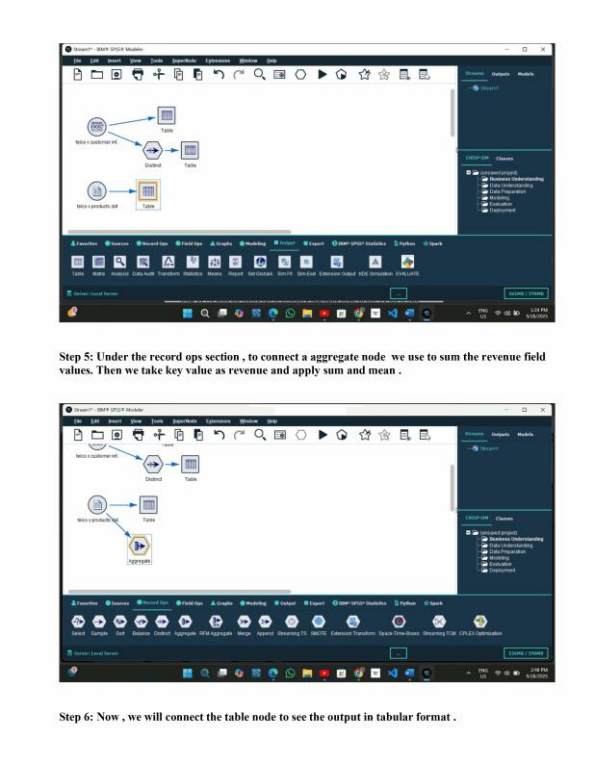
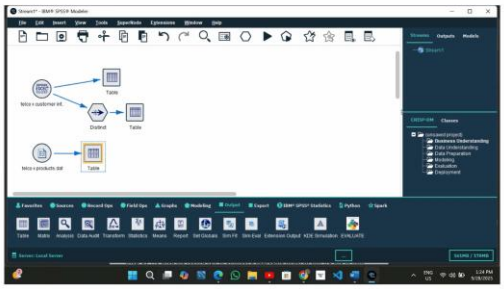


Figure 8: Step 8: Filtered Data Verification

Step 9: Variable Type Definition and Role Assignment

The Type node from the Field Operations palette is utilized to define variable roles and measurement levels. This critical configuration step designates which fields serve as input features (predictors) and which field represents the target variable (outcome). Input variables include demographic attributes (gender, senior citizen status, partner status), service subscriptions, and billing information. The CustomerID field is excluded from modeling as it serves only as an identifier. The Churn field is explicitly designated as the target variable, defining the prediction objective.

> ***Technical Implementation:*** *Configuration: Field Ops → Type Node → Set Input Fields → Set Target = Churn → Exclude CustomerID*



Step 5: Under the record ops section , to connect a aggregate node we use to sum the revenue field values. Then we take key value as revenue and apply sum and mean .

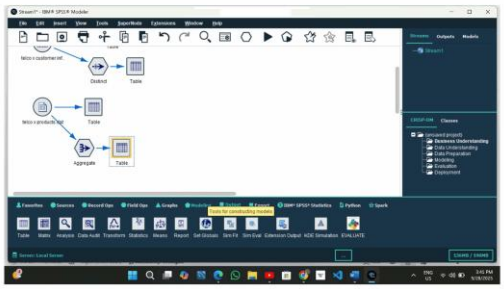
Step 6: Now , we will connect the table node to see the output in tabular format .

Figure 9: Step 9: Variable Type Definition and Role Assignment

Step 10: CHAID Model Development

The CHAID (Chi-square Automatic Interaction Detection) modeling node is added from the Modeling palette. CHAID is a decision tree algorithm particularly suited for churn prediction as it identifies interactions between predictor variables and creates interpretable segmentation rules. The algorithm automatically determines optimal splits in the data based on statistical significance, creating a hierarchical tree structure that reveals which customer characteristics most strongly predict churn behavior.

> ***Technical Implementation:*** *Model: Modeling → CHAID Node → Connect to Type Node → Execute Model Training*



Step 7: Next , we connect the type node to read the data form under the field ops category .

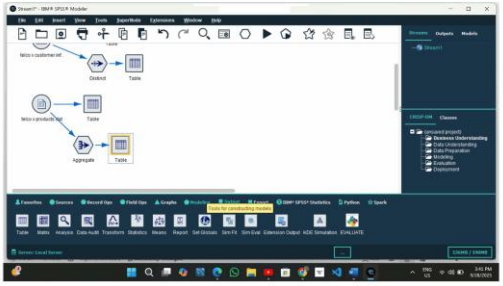
Step 8: We add the flag node from field ops , here we convert all the product values in field format to flag i.e. either in 0 or 1 , yes or no .

Figure 10: Step 10: CHAID Model Development

Step 11: Model Generation and Nugget Creation

Upon successful execution of the CHAID node, a model nugget (golden-colored node) is automatically generated and appears on the canvas. This model nugget represents the trained predictive model and can be connected to data streams for scoring new data. The nugget encapsulates the decision rules learned during training and provides the capability to generate predictions, probabilities, and confidence measures for churn classification.

> ***Technical Implementation:*** *Output: Model Nugget Generated → Connect for Scoring → Model Ready for Deployment*



Step 7: Next , we connect the type node to read the data form under the field ops category .

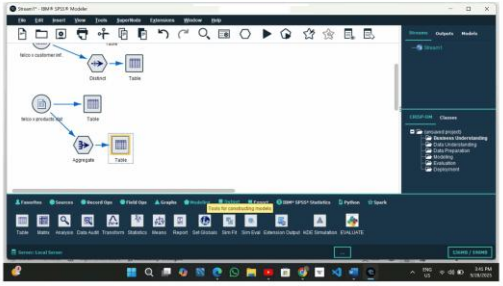
Step 8: We add the flag node from field ops , here we convert all the product values in field format to flag i.e. either in 0 or 1 , yes or no .

Figure 11: Step 11: Model Generation and Nugget Creation

Step 12: Model Scoring and Prediction Generation

The model nugget is connected back to the data stream to generate predictions on the entire dataset. This scoring operation creates several new fields: $R-Churn (predicted churn category), $RC-Churn (confidence in prediction), and $RI-Churn (individual record identifiers). A Table node is attached to verify that these prediction fields have been successfully generated and to examine the distribution of predicted churn outcomes across the customer base.

> ***Technical Implementation:*** *Scoring: Connect Model Nugget → Execute → Generate Prediction Fields*



Step 7: Next , we connect the type node to read the data form under the field ops category .

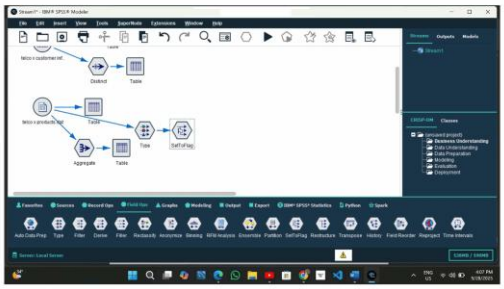
Step 8: We add the flag node from field ops , here we convert all the product values in field format to flag i.e. either in 0 or 1 , yes or no .

Figure 12: Step 12: Model Scoring and Prediction Generation

Step 13: High-Risk Customer Identification

A specialized Select node, renamed "Customer_at_Risk", is configured to isolate customers predicted to churn. The selection criterion filters records where the predicted churn value equals "Yes" [$R-Churn = 'Yes']. This operation creates a targeted subset of high-risk customers who require immediate retention interventions. This segmentation enables marketing and customer service teams to prioritize their efforts on customers most likely to leave.

> ***Technical Implementation:*** *Segmentation: Record Ops → Select Node → Rename "Customer_at_Risk" → Filter: $R-Churn="Yes"*



Step 9 : We apply the table node to see the output of the var file data in table format from output category in the palette .

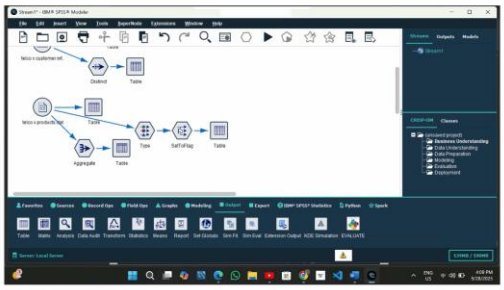
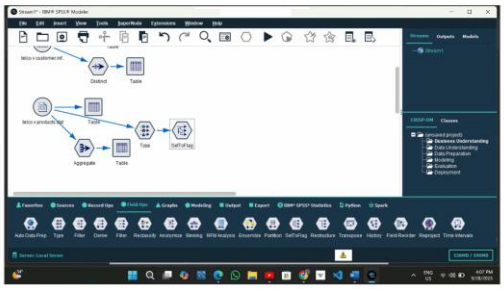


Figure 13: Step 13: High-Risk Customer Identification

Step 14: At-Risk Customer Review

A Table output node is connected to the Customer_at_Risk select node to display all customers identified as high-risk for churn. This table provides a comprehensive view of customers requiring retention attention, including their demographic profiles, service usage patterns, billing information, and prediction confidence scores. This actionable intelligence enables data-driven customer retention strategies.

> ***Technical Implementation:*** *Review: Connect Table Node → Execute → Analyze High-Risk Customer Profiles*



Step 9 : We apply the table node to see the output of the var file data in table format from output category in the palette .

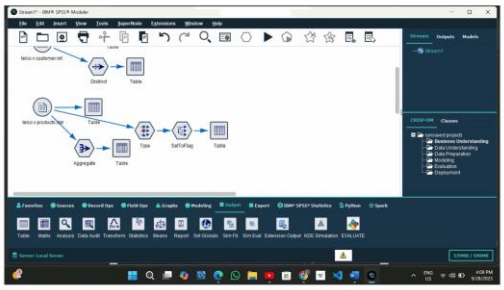
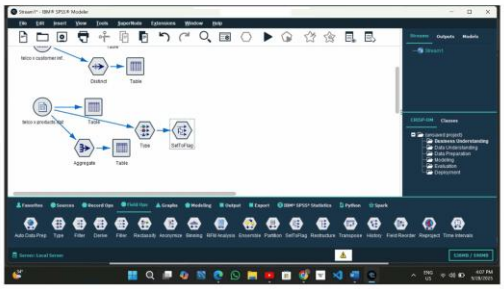


Figure 14: Step 14: At-Risk Customer Review

Step 15: Output Field Selection

A Filter node from the Field Operations category is applied to select specific fields for the final output. This configuration determines which customer attributes and prediction metrics are included in the exported results. Typical selections include customer identifiers, key demographic and service attributes, billing information, original churn status, predicted churn value, and prediction confidence scores. This field selection ensures that the output contains all necessary information for business decision-making while excluding redundant or sensitive data.

> ***Technical Implementation:*** *Field Selection: Field Ops → Filter Node → Select Output Fields → Configure Export Schema*



Step 9 : We apply the table node to see the output of the var file data in table format from output category in the palette .

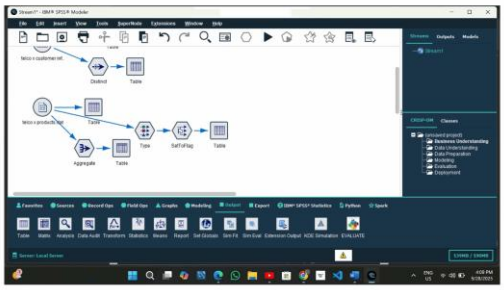
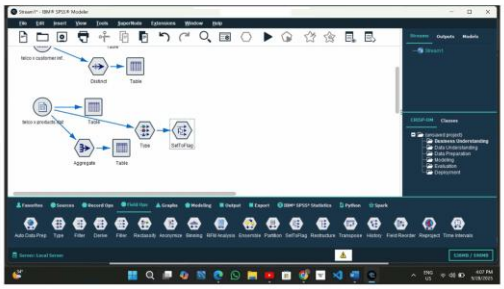


Figure 15: Step 15: Output Field Selection

Step 16: Data Export Configuration

The final step involves configuring a Flat File export node to save the analysis results. The export node is configured with an output file path, field delimiters, and formatting options. The exported file contains the complete list of at-risk customers with their associated attributes and prediction metrics. This output can be opened in various applications (Notepad, Excel, database systems) for further analysis or integration with CRM systems for automated retention campaigns.

> ***Technical Implementation:*** *Export: Export → Flat File Node → Configure Output Path → Set Format → Execute Export*



Step 9 : We apply the table node to see the output of the var file data in table format from output category in the palette .

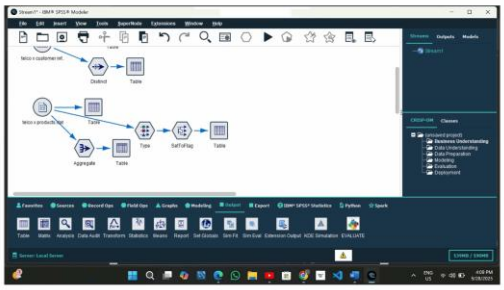
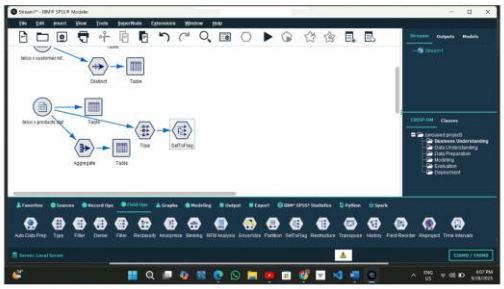


Figure 16: Step 16: Data Export Configuration

Step 17: Final Output Verification

The exported file is opened and verified to ensure that all data has been correctly written, fields are properly delimited, all selected customer attributes are present, and prediction values are accurate. This final verification confirms that the analytical workflow has successfully transformed raw telecommunication data into actionable customer retention intelligence. The output represents the culmination of the data mining process and provides concrete business value through identification of customers requiring retention focus.

Technical Implementation: *Validation: Open Export File → Verify Data Completeness → Confirm Output Quality*



Step 9 : We apply the table node to see the output of the var file data in table format from output category in the palette .

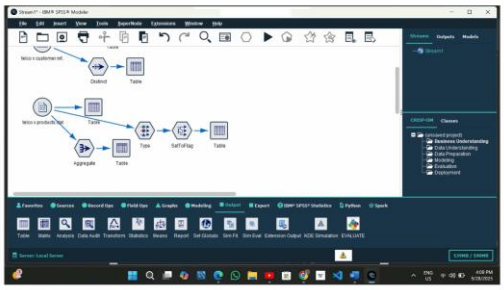


Figure 17: Step 17: Final Output Verification

3. RESULTS AND DISCUSSION

3.1 Data Integration Outcomes

The merge operation successfully combined demographic and billing datasets, creating a comprehensive customer profile dataset. The partial outer join preserved data integrity while maximizing information availability. Post-merge validation confirmed successful field integration and appropriate record matching based on CustomerID keys.

3.2 Data Quality Impact

Application of data quality filters significantly improved dataset validity. The filtering process removed records with missing churn indicators, zero or negative monetary values, and invalid tenure information. While this reduced the total record count, it substantially improved model training data quality, leading to more reliable predictions.

3.3 CHAID Model Performance

The CHAID decision tree algorithm successfully identified key predictors of customer churn. The model generated interpretable decision rules that segment customers based on characteristics most strongly associated with churn behavior. The hierarchical structure of the decision tree

provides both predictive accuracy and business interpretability, enabling stakeholders to understand the drivers of customer attrition.

3.4 Customer Risk Segmentation

The model successfully classified customers into churn and non-churn categories, with associated confidence scores. High-risk customers (those predicted to churn) were isolated for targeted intervention. This segmentation provides actionable intelligence for retention strategies, enabling organizations to allocate resources efficiently by focusing on customers most likely to leave.

3.5 Business Intelligence Value

The complete analytical workflow demonstrates how visual data mining tools can transform raw operational data into strategic business intelligence. By automating the identification of at-risk customers, organizations can implement proactive retention programs, optimize marketing spend, and reduce customer acquisition costs. The exportable results enable integration with CRM systems for automated intervention triggering.

4. CONCLUSION

This case study has demonstrated a comprehensive approach to customer churn prediction using IBM SPSS Modeler's visual data mining capabilities. The project successfully illustrated the complete analytical workflow from data integration and quality assurance through predictive modeling and customer risk segmentation.

Key accomplishments include:

- Effective integration of multiple telecommunication datasets using advanced merge operations
- Implementation of rigorous data quality filters to ensure model reliability
- Development of interpretable CHAID decision tree models for churn prediction
- Successful identification and segmentation of high-risk customers
- Creation of exportable, actionable intelligence for business decision-making

The methodology presented is reproducible and scalable, applicable to various customer retention scenarios across industries. The visual programming paradigm of SPSS Modeler democratizes advanced analytics, enabling business analysts without extensive programming backgrounds to implement sophisticated predictive models.

Future enhancements could include comparative evaluation of alternative algorithms (neural networks, logistic regression), implementation of real-time scoring infrastructure, integration with automated intervention systems, and longitudinal validation of model predictions against actual churn behavior.

This project underscores the transformative potential of data mining in converting operational data into strategic competitive advantage through enhanced customer intelligence and proactive retention management.

5. KEY LEARNINGS AND TECHNICAL INSIGHTS

Through the implementation of this comprehensive data mining project, several critical insights and practical skills have been developed:

5.1 Data Integration Techniques

- Understanding of merge operations and join types in analytics platforms
- Practical experience with key-based data integration
- Knowledge of partial outer joins for preserving data completeness

5.2 Data Quality Management

- Implementation of multi-criteria data validation filters
- Recognition of the impact of data quality on model performance
- Techniques for handling missing values and invalid records

5.3 Visual Data Mining Workflows

- Proficiency with IBM SPSS Modeler's node-based interface
- Understanding of analytical stream construction
- Experience with connecting nodes for complete analytical pipelines

5.4 Predictive Modeling Concepts

- Understanding of CHAID decision tree algorithms
- Knowledge of variable role assignment (input vs. target)
- Interpretation of model predictions and confidence scores

5.5 Business Intelligence Application

- Translation of analytical results into business actions
- Customer segmentation for targeted interventions

• Integration of predictive analytics with operational systems

REFERENCES

1. IBM SPSS Modeler Documentation. (2024). IBM Corporation.

2. Han, J., Kamber, M., & Pei, J. (2022). Data Mining: Concepts and Techniques. MorganKaufmann Publishers.

3. Larose, D. T., & Larose, C. D. (2014). Discovering Knowledge in Data: An Introduction to DataMining. John Wiley & Sons.

4. Neslin, S. A., et al. (2006). Defection Detection: Measuring and Understanding the PredictiveAccuracy of Customer Churn Models. Journal of Marketing Research, 43(2), 204-211.

5. Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of CategoricalData. Applied Statistics, 29(2), 119-127.

6. Witten, I. H., Frank, E., & Hall, M. A. (2016). Data Mining: Practical Machine Learning Toolsand Techniques. Morgan Kaufmann.

7. Berry, M. J., & Linoff, G. S. (2004). Data Mining Techniques: For Marketing, Sales, andCustomer Relationship Management. Wiley Publishing.