MASTER OF DATA SCIENCE (SEMESTER 1 – 2022/2023)

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

WQD7005 DATA MINING

GROUP PROJECT

Analyzing Customer Churn In The Telecommunications Industry

GROUP MEMBERS	
22081192	TNEU ZHEN YANG
22062358	HAN DIE
22060253	SIYU JIANG

INSTRUCTOR: PROF DR TEH YING WAH
DATE: 16th JANUARY 2024

Contents

Chapter 1: Introduction **3**
Chapter 2: Business Understanding **4**
2.1 Analysis Goal 4
2.2 Analysis Data 4
2.3 Objectives 5
Chapter 3: Dataset Description **5**
Chapter 4: SEMMA Process **13**
4.1 SEMMA Description 13
4.2 Data Integration 14
4.3 Sample (SEMMA: Step 1) 25
4.4 Explore (SEMMA: Step 2) 26
4.4.1 Importing and Summary Statistics 26
4.4.2 Univariate Analysis 29
4.4.3 Bivariate Analysis 32
4.4.4 Trivariate Analysis 36
4.4.5 Multivariate Analysis 38
4.4.6 Clustering Analysis 40
4.5 Modify (SEMMA: Step 3) 43
4.5.1 Fixing Misspelling and Data Errors 43
4.5.2 Missing Value Imputation 46
4.5.3 Feature Selection 48
4.5.4 Normalization 48
4.6 Model (SEMMA: Step 4) 49
4.6.1 Importing Data 49
4.6.2 Data Partition 54
4.6.3 Decision Tree Model 55
4.6.4 Gradient Boosting Model 56
4.6.5 Neural Network Model 56
4.7 Assess (SEMMA: Step 5) 57
4.7.1 Model Comparison 58
4.7.2 Variable Importance Analysis 61
4.7.2 Oversampling 63
Chapter 5: Conclusion and Recommendations **64**
5.1 Summary of Insights 64
5.2 Recommendations 65
5.3 Reflections 66
Appendix **67**

Chapter 1: Introduction

Customer churn, the loss of customers or subscribers, is a critical facet that can significantly impact a company's profitability. For businesses, the cost-effectiveness of retaining existing customers over acquiring new ones underscores the importance of understanding and mitigating customer churn. In the telecommunications industry, customer churn is especially pivotal, serving as a compass that guides companies through the complexities of customer retention.

In the context of the Telco customer churn dataset, unraveling the intricacies of customer churn is paramount. This dataset, comprising demographics, satisfaction scores, churn scores, and customer lifetime value indices, offers a comprehensive lens for companies to explore factors influencing customer departure. Beyond its direct impact on revenue, customer churn serves as a reflection of customer satisfaction and the overall health of the business. A proactive approach to identifying these factors empowers companies to address customer needs, elevate service quality, and establish robust strategies for fostering long-term customer loyalty.

The Telco customer churn dataset, sourced from IBM, encompasses a wealth of information critical for understanding and addressing customer churn in the telecommunications sector. This dataset's significance lies in its ability to provide a holistic view of customer interactions, ranging from demographics to satisfaction metrics. It allows for a detailed exploration of patterns and correlations that underlie customer behavior, enabling data-driven strategies to enhance customer satisfaction and reduce churn.

Drawing a parallel with the example report on store sales data, the Telco customer churn dataset shares a common objective in leveraging data science for decision-making. While the store sales data focuses on time series forecasting to manage inventory and increase profitability, the Telco customer churn dataset aims to uncover insights for optimizing

customer retention strategies. Both datasets showcase the relevance of data science in addressing real-world challenges within their respective industries.

As we delve into the Telco customer churn dataset, we aim to enhance customer satisfaction, reduce churn, and ultimately contribute to the sustainable growth of the telecommunications business.

Chapter 2: Business Understanding

2.1 Analysis Goal

The primary goal is to understand and analyze customer churn within the telco company, as represented by the Telco customer churn dataset. Customer churn, the phenomenon of customers discontinuing services, holds significant implications for business profitability and customer satisfaction. The overarching goal is to identify patterns, correlations, and predictive indicators within the dataset that can inform strategic decision-making to mitigate churn and enhance customer retention.

2.2 Analysis Data

The Telco customer churn dataset comprises a wealth of information encompassing customer demographics, location, population estimates, services subscribed, and status metrics such as Satisfaction Score, Churn Score, and Customer Lifetime Value (CLTV). The dataset's richness provides a holistic view of customer interactions, allowing for a multifaceted analysis of the factors influencing churn. By leveraging this diverse set of

data, the analysis aims to uncover the key drivers of customer attrition and formulate data-driven strategies for reducing churn rates.

2.3 Objectives

The objectives of this data mining project are as follows.

1. To explore the relationship between the variables and the target variable, customer churn.
2. To build a model to predict customer churn using selected features.
3. To assess the model using reasonable metrics.

Chapter 3: Dataset Description

The Telco customer churn data contains information about a fictional telco company that provided home phone and Internet services to 7043 customers in California in Q3. It indicates which customers have left, stayed, or signed up for their service. Multiple important demographics are included for each customer, as well as a Satisfaction Score, Churn Score, and Customer Lifetime Value (CLTV) index.

The Telco customer churn dataset, sourced from IBM, and the Kaggle Store Sales dataset differ in focus and scope. The Telco dataset, with its five files, explores customer demographics and services in a California-based telco company during Q3, aiming to understand and reduce customer churn. In contrast, the Kaggle Store Sales dataset concentrates on time series forecasting for grocery stores in Ecuador, emphasizing inventory management and profitability. While both datasets highlight the versatility of data science, they cater to unique industry needs – telecommunications customer retention and retail sales forecasting, respectively.

The Telco_customer_churn_demographics data delves into the demographic details of each customer, providing a snapshot of their gender, age, marital status, presence of dependents, and more. This rich demographic information is pivotal for understanding the varied customer profiles that contribute to the overall churn landscape. Table 2.1 shows the variables' description.

Table 2.1 Telco_customer_churn_demographics

Variable	**Description**
Customer ID	A unique identifier assigned to each customer
Count	A value used in reporting/dashboarding to sum up the number of customers in a filtered set
Gender	The customer's gender: Male, Female
Age	The customer's current age, in years, at the time the fiscal quarter ended.
Under 30	Indicates whether the customer is under 30 years old: Yes, No
Senior Citizen	Indicates if the customer is 65 or older: Yes, No
Married	Indicates if the customer is married: Yes, No
Dependents	Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.
Number of Dependents	Indicates the number of dependents that live with the customer.

Geographical insights play a crucial role in the telecommunications landscape. The Telco_customer_churn_location data offers a detailed perspective on the distribution of customers across states, cities, and zip codes in California. Analyzing location-specific churn patterns can uncover regional dynamics and aid in targeted business strategies. Table 2.2 shows the variables' description.

Table 2.2 Telco_customer_churn_location

Variable	**Description**
Location ID	A unique identifier assigned to each location
Customer ID	A unique identifier assigned to each customer
Count	A value used in reporting/dashboarding to sum up the number of customers in a filtered set
Country	The country of the customer's primary residence
State	The state of the customer's primary residence
City	The city of the customer's primary residence
Zip Code	The zip code of the customer's primary residence
Lat Long	The combined latitude and longitude of the customer's primary residence
Latitude	The latitude of the customer's primary residence

Longitude	The longitude of the customer's primary residence

Understanding the population dynamics within specific zip code areas is vital for contextualizing customer behavior. The Telco_customer_churn_population data provides estimates of the population in each zip code, enriching the dataset with demographic context that can influence churn trends. Table 2.3 shows the variables' description.

Table 2.3 Telco_customer_churn_population

Variable	**Description**
ID	A unique identifier assigned to each row in the population dataset
Zip Code	The zip code of the customer's primary residence
Population	A current population estimate for the entire zip code area

The Telco_customer_churn_services data captures intricate details about the services subscribed to by customers. From referral patterns and tenure to specific service offerings like internet type and contract details, this data offers a comprehensive overview of the services influencing customer satisfaction and, consequently, churn. Table 2.4 shows the variables' description.

Table 2.4 Telco_customer_churn_services.xlsx

Variable	**Description**
Service ID	A unique identifier assigned to each service entry

Customer ID	A unique identifier assigned to each customer
Count	A value used in reporting/dashboarding to sum up the number of customers in a filtered set
Quarter	The fiscal quarter that the data has been derived from (e.g., Q3)
Referred a Friend	Indicates if the customer has ever referred a friend or family member to this company (Yes/No)
Number of Referrals	Indicates the number of referrals to date that the customer has made
Tenure in Months	Indicates the total amount of months that the customer has been with the company by the end of the specified quarter
Offer	Identifies the last marketing offer that the customer accepted, if applicable (e.g., Offer A, Offer B, Offer C, Offer D, Offer E)
Phone Service	Indicates if the customer subscribes to home phone service with the company (Yes/No)
Avg Monthly Long Distance Charges	Indicates the customer's average long-distance charges, calculated to the end of the specified quarter
Multiple Lines	Indicates if the customer subscribes to multiple telephone lines with the company (Yes/No)

Internet Service	Indicates if the customer subscribes to Internet service with the company (Yes/No)
Internet Type	Specifies the type of Internet service subscribed (e.g., DSL, Fiber Optic Cable)
Avg Monthly GB Download	Indicates the customer's average download volume in gigabytes, calculated to the end of the specified quarter
Online Security	Indicates if the customer subscribes to an additional online security service provided by the company (Yes/No)
Online Backup	Indicates if the customer subscribes to an additional online backup service provided by the company (Yes/No)
Device Protection Plan	Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company (Yes/No)
Premium Tech Support	Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times (Yes/No)
Streaming TV	Indicates if the customer uses their Internet service to stream television programming from a third-party provider (Yes/No)
Streaming Movies	Indicates if the customer uses their Internet service to stream movies from a third-party provider (Yes/No)
Streaming Music	Indicates if the customer uses their Internet service to stream music from a third-party provider (Yes/No)

Unlimited Data	Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads (Yes/No)
Contract	Indicates the customer's current contract type (Month-to-Month, One Year, Two Year)
Paperless Billing	Indicates if the customer has chosen paperless billing (Yes/No)
Payment Method	Indicates how the customer pays their bill (Bank Withdrawal, Credit Card, Mailed Check)
Monthly Charge	Indicates the customer's current total monthly charge for all their services from the company
Total Charges	Indicates the customer's total charges, calculated to the end of the specified quarter
Total Refunds	Indicates the customer's total refunds, calculated to the end of the specified quarter
Total Extra Data Charges	Indicates the customer's total charges for extra data downloads above those specified in their plan, by the end of the specified quarter
Total Long Distance Charges	Indicates the customer's total charges for long-distance above those specified in their plan, by the end of the specified quarter
Total Revenue	The total revenue generated from the customer, including charges, refunds, extra data charges, and long-distance charges

At the heart of the dataset is the Telco_customer_churn_status, encapsulating the status of each customer at the end of Q3. With information on satisfaction scores, churn labels, churn values, churn scores, and Customer Lifetime Value (CLTV), this file is a treasure trove for predicting, understanding, and mitigating churn. Table 2.5 shows the variables' description.

Table 2.5 Telco_customer_churn_status

Variable	**Description**
Status ID	A unique identifier assigned to each status entry
Customer ID	A unique identifier assigned to each customer
Count	A value used in reporting/dashboarding to sum up the number of customers in a filtered set
Quarter	The fiscal quarter that the data has been derived from (e.g., Q3)
Satisfaction Score	A customer's overall satisfaction rating of the company from 1 (Very Unsatisfied) to 5 (Very Satisfied)
Customer Status	Indicates the status of the customer at the end of the specified quarter (Churned, Stayed, or Joined)
Churn Label	Yes = the customer left the company this quarter, No = the customer remained with the company

Churn Value	1 = the customer left the company this quarter, 0 = the customer remained with the company
Churn Score	A value from 0-100 that is calculated using the predictive tool IBM SPSS Modeler. The higher the score, the more likely the customer will churn
CLTV (Customer Lifetime Value)	A predicted CLTV calculated using corporate formulas and existing data. The higher the value, the more valuable the customer. High-value customers should be monitored for churn
Churn Category	A high-level category for the customer's reason for churning (e.g., Attitude, Competitor, Dissatisfaction, Price)
Churn Reason	A customer's specific reason for leaving the company, directly related to Churn Category

Chapter 4: SEMMA Process

4.1 SEMMA Description

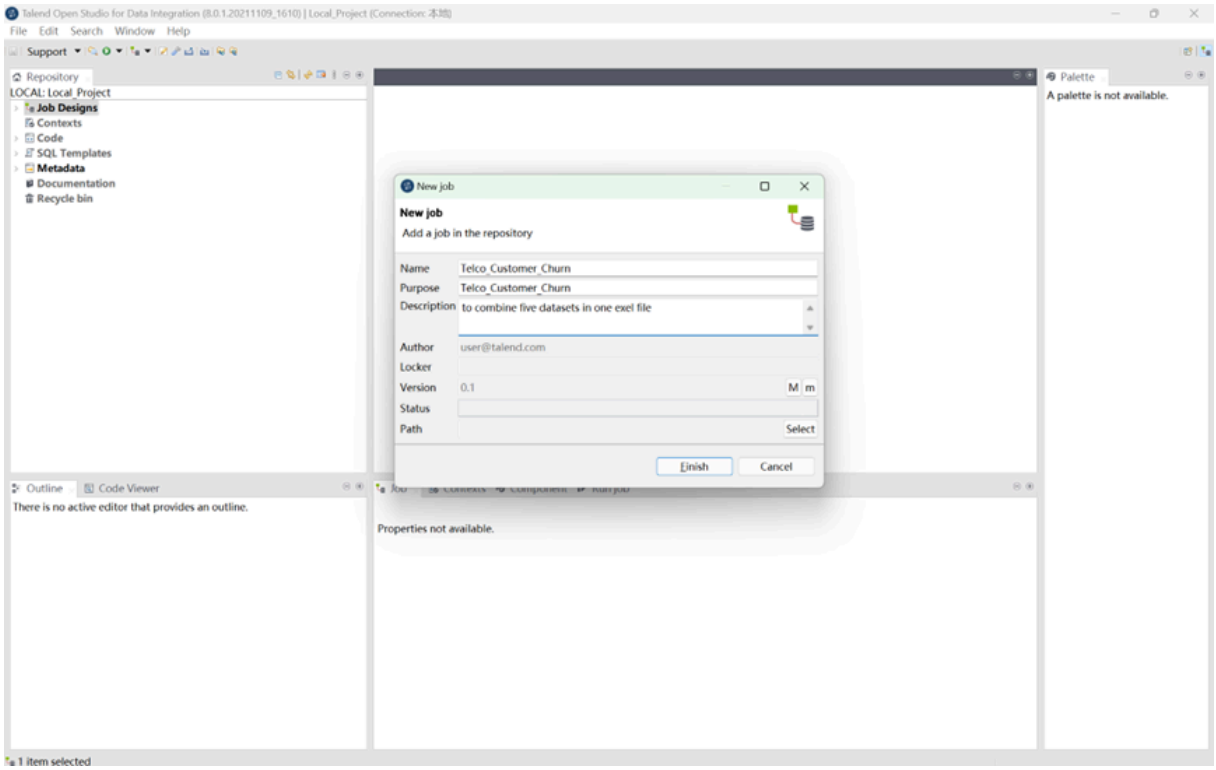
SEMMA is an acronym representing a structured and systematic data mining process that stands for Sample, Explore, Modify, Model, and Assess. This methodology is widely used in the field of data mining and analytics to guide the end-to-end process of extracting valuable insights from data. The SEMMA process provides a systematic framework for data scientists and analysts to navigate the complexities of data mining projects. It emphasizes a structured approach, starting from data sampling, through

exploration and modification, to modeling and final assessment, ensuring a comprehensive and reliable analysis of the dataset. This approach is particularly relevant when applied to The Telco customer churn dataset, offering a step-by-step framework to extract valuable insights.

4.2 Data Integration

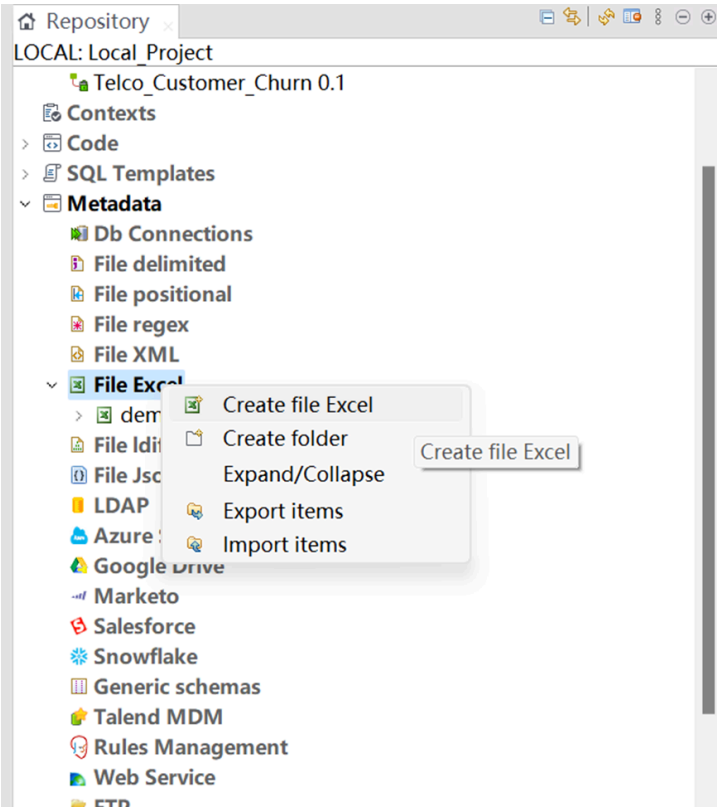
The Telco customer churn datasets are including five excel files which are Telco_customer_churn_demographics,Telco_customer_churn_location,Telco_customer_churn_population, Telco_customer_churn_services, and Telco_customer_churn_status. So we need to combine these five excel files into one excel file, as the input dataset of the SEMMA process. We will use Talend, an open-source software integration platform that provides a comprehensive suite of tools for various data integration and management tasks, to extract the data from above five excel files, combine (transform) all the data together, and load the combined data into a new excel file. Here are the steps in Talend:

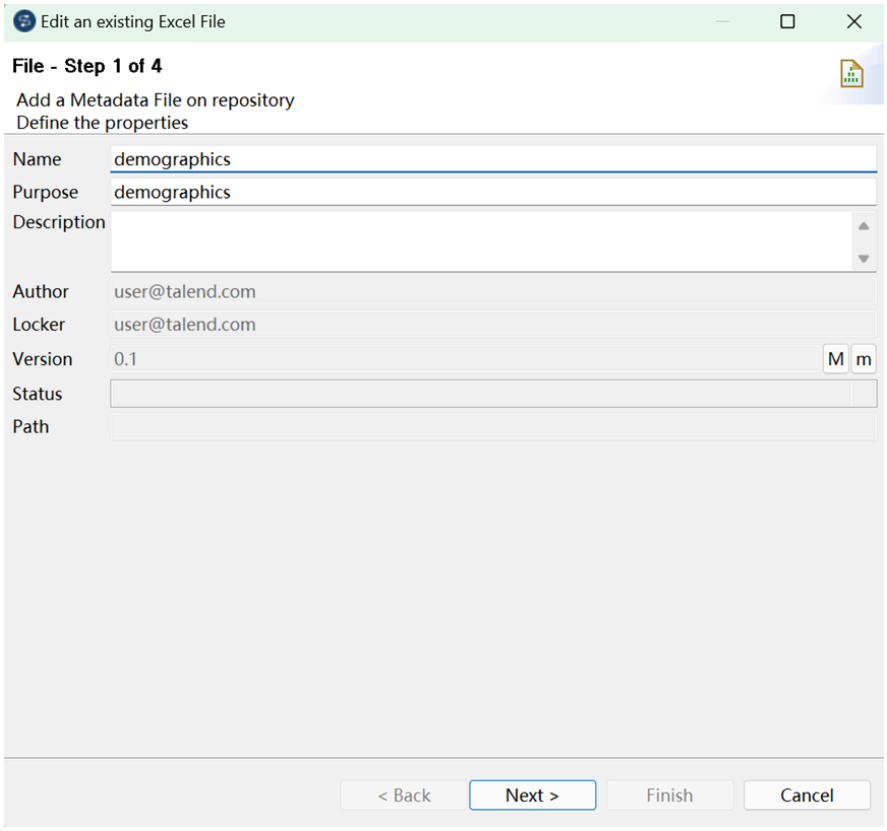
1. Create a new job with name and description of the job:

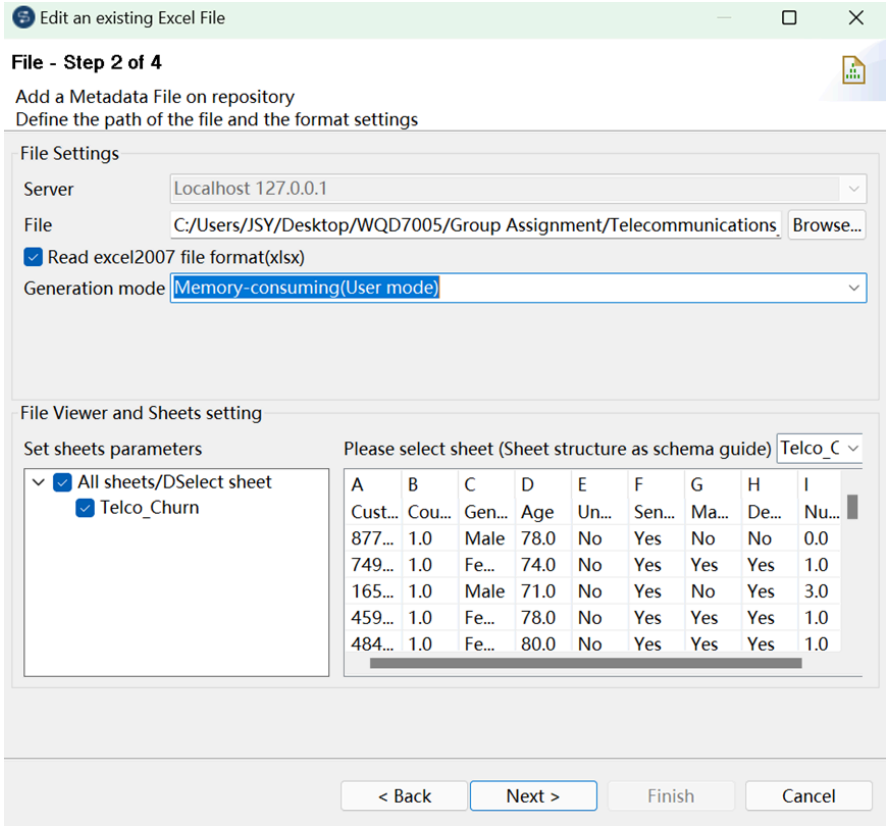


2. Add a tFileInputExcel component to input the data from Telco_customer_churn_demographics:

3. Add a metadata file on repository and configure the schema:


Repository
LOCAL: Local_Project
Telco_Customer_Churn 0.1
Contexts
Code
SQL Templates
Metadata
Db Connections
File delimited
File positional
File regex
File XML
Create file Excel
Create folder
Expand/Collapse
Export items
Import items
Create file Excel
LDAP
Marketo
Salesforce
Snowflake
Generic schemas
Talend MDM
Rules Management
Web Service


Edit an existing Excel File
File - Step 1 of 4
Add a Metadata File on repository
Define the properties
Name
demographics
Purpose
demographics
Description
Author
user@talend.com
Locker
user@talend.com
Version
0.1
M
m
Status
Path
< Back
Next >
Finish
Cancel


Edit an existing Excel File
File - Step 2 of 4
Add a Metadata File on repository
Define the path of the file and the format settings
File Settings
Server
Localhost 127.0.0.1
File
C:/Users/JSY/Desktop/WQD7005/Group Assignment/Telecommunications
Browse...
Read excel2007 file format(xlsx)
Generation mode
Memory-consuming(User mode)
File Viewer and Sheets setting
Set sheets parameters
All sheets/DSelect sheet
Telco_Churn
Please select sheet (Sheet structure as schema guide)
Telco_C
A B C D E F G H I
Cust... Cou... Gen... Age Un... Sen... Ma... De... Nu...
877... 1.0 Male 78.0 No Yes No No 0.0
749... 1.0 Fe... 74.0 No Yes Yes Yes 1.0
165... 1.0 Male 71.0 No Yes No Yes 3.0
459... 1.0 Fe... 78.0 No Yes Yes Yes 1.0
484... 1.0 Fe... 80.0 No Yes Yes Yes 1.0
< Back
Next >
Finish
Cancel

Edit an existing Excel File

File - Step 3 of 4

Add a Metadata File on repository
Define the setting of the parse job

File Settings
Encoding UTF-8
Advanced separator(for number)
Thousands separator: ','
Decimal separator: '.'

Rows To Skip
If any rows must be ignored, specify the following parameters
Header 1
Footer

Metadata column setting
First column: 1
Last column:

Limit Of Rows
If the number of lines must be limited, specify this number
Limit

Preview Output

Set heading row as column names Refresh Preview

Customer ID	Count	Gender	Age	Under 30	Senior Citizen	Married	Dependents	Number of
8779-QRDMV	1	Male	78	No	Yes	No	No	0
7495-OOKFY	1	Female	74	No	Yes	Yes	Yes	1
1658-BYGOY	1	Male	71	No	Yes	No	Yes	3

Export as context Revert Context

< Back Next > Finish Cancel

Schema

Update Schema "demographics"

Update an existing Schema in repository

Name demographics
Comment

Schema

Click Guess button to update the schema below according to your settings

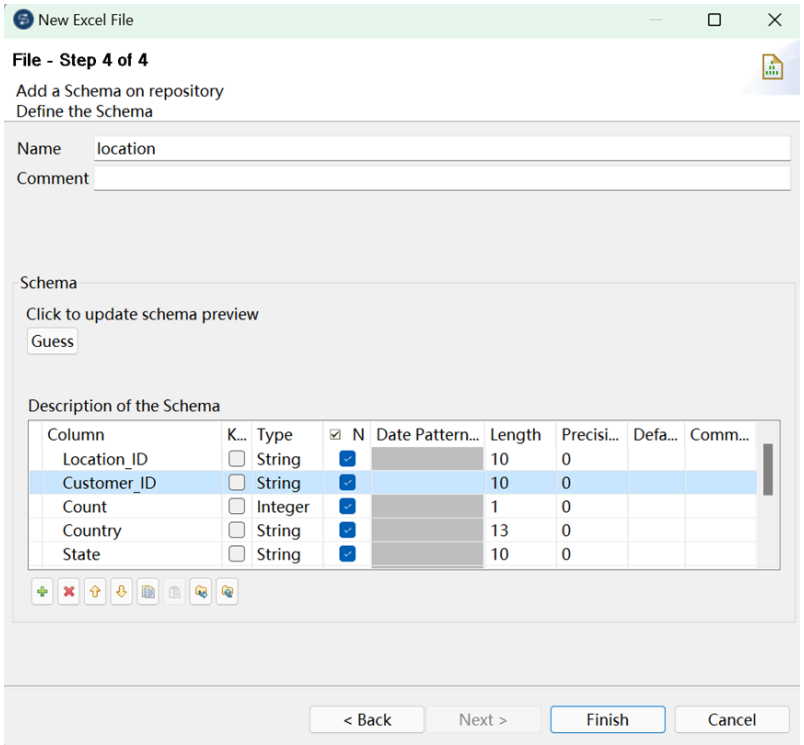
Guess

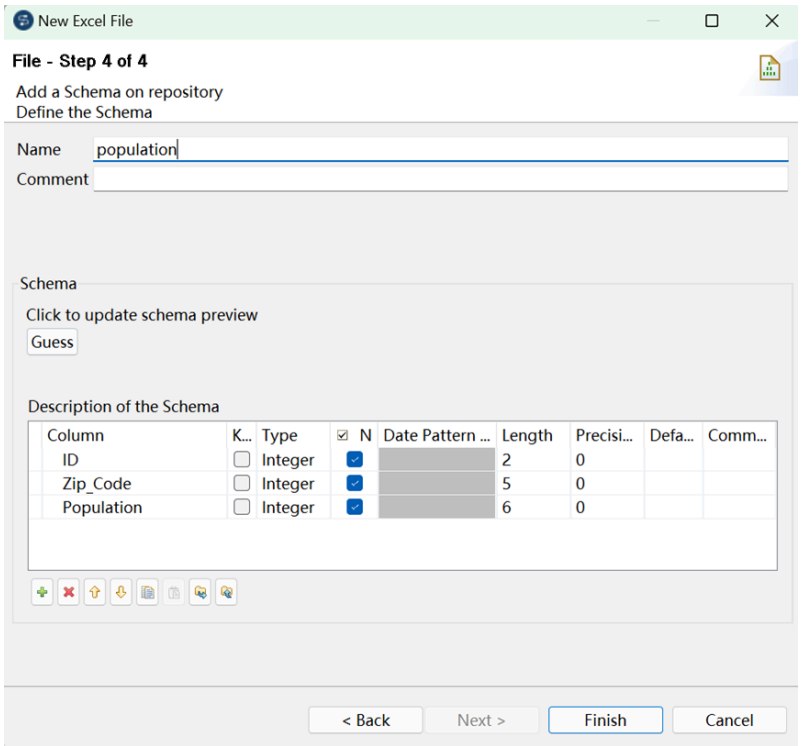
Description of the Schema

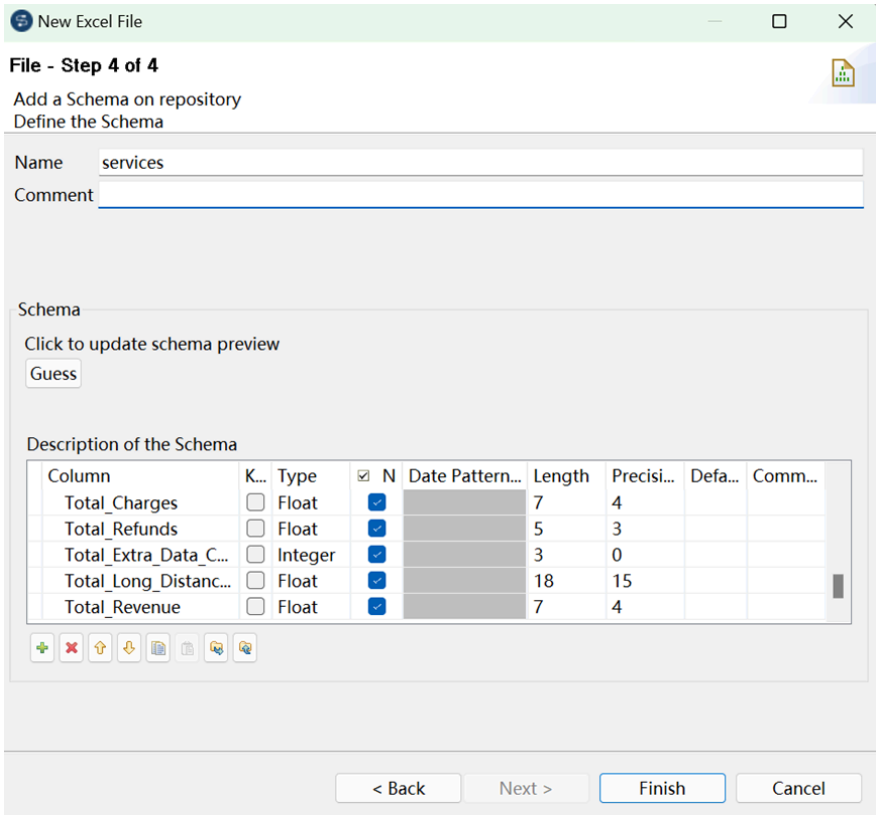
Column	K...	Type	N	Date Patter...	Length	Precisi...	Def...	Com...
Customer_ID		String			10	0		
Count		Integer			1	0		
Gender		String			6	0		
Age		Integer			2	0		

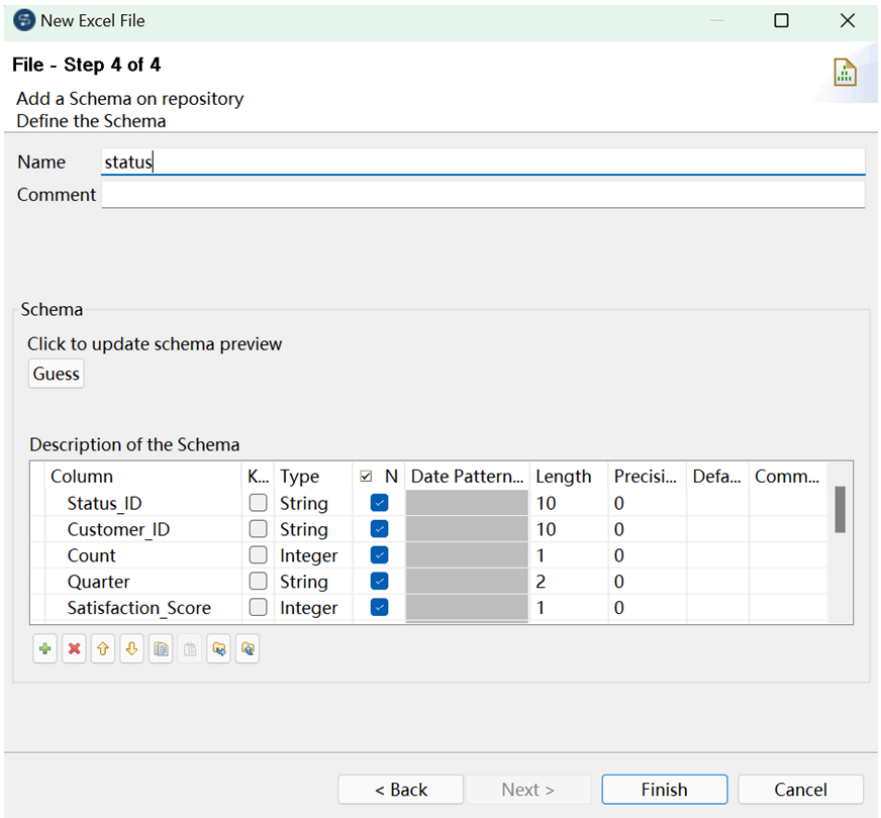
Finish Cancel

4. Input data from Telco_customer_churn_location, Telco_customer_churn_population, Telco_customer_churn_services, and Telco_customer_churn_status:

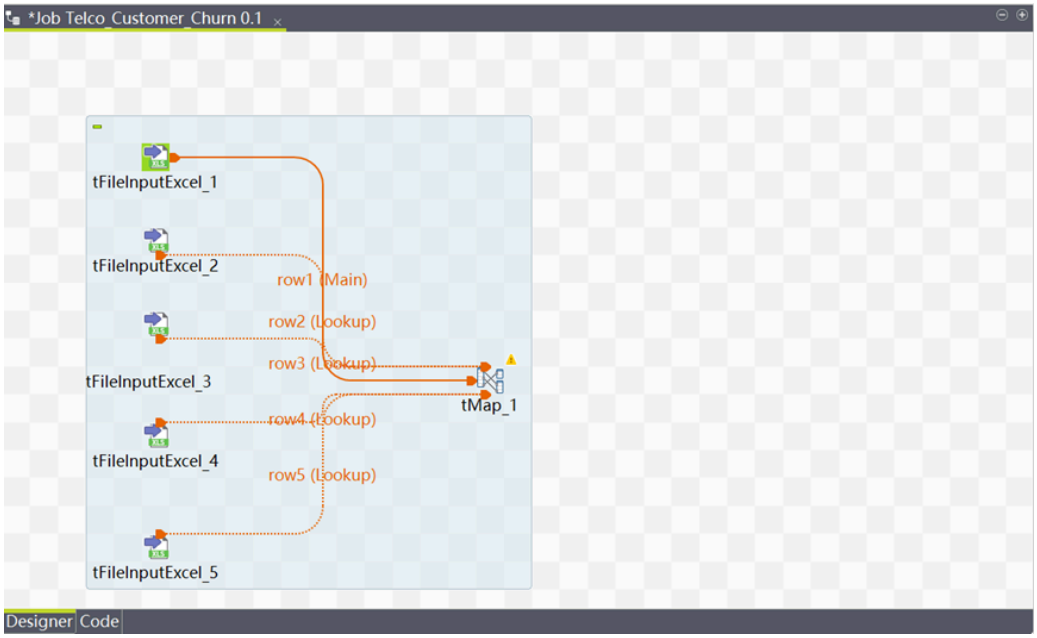




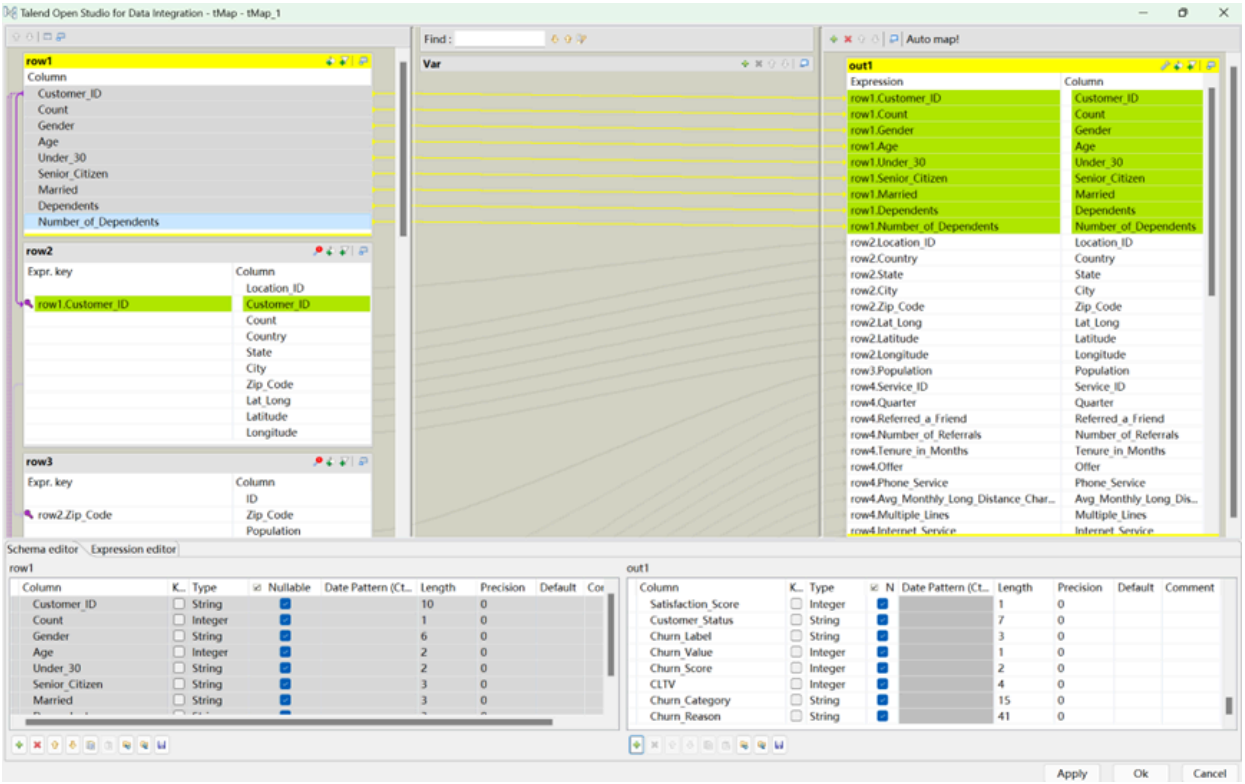

New Excel File
File - Step 4 of 4
Add a Schema on repository
Define the Schema
Name services
Comment
Schema
Click to update schema preview
Guess
Description of the Schema
Column K... Type N Date Pattern... Length Precisi... Defa... Comm...
Total_Charges Float 7 4
Total_Refunds Float 5 3
Total_Extra_Data_C... Integer 3 0
Total_Long_Distanc... Float 18 15
Total_Revenue Float 7 4
< Back
Next >
Finish
Cancel

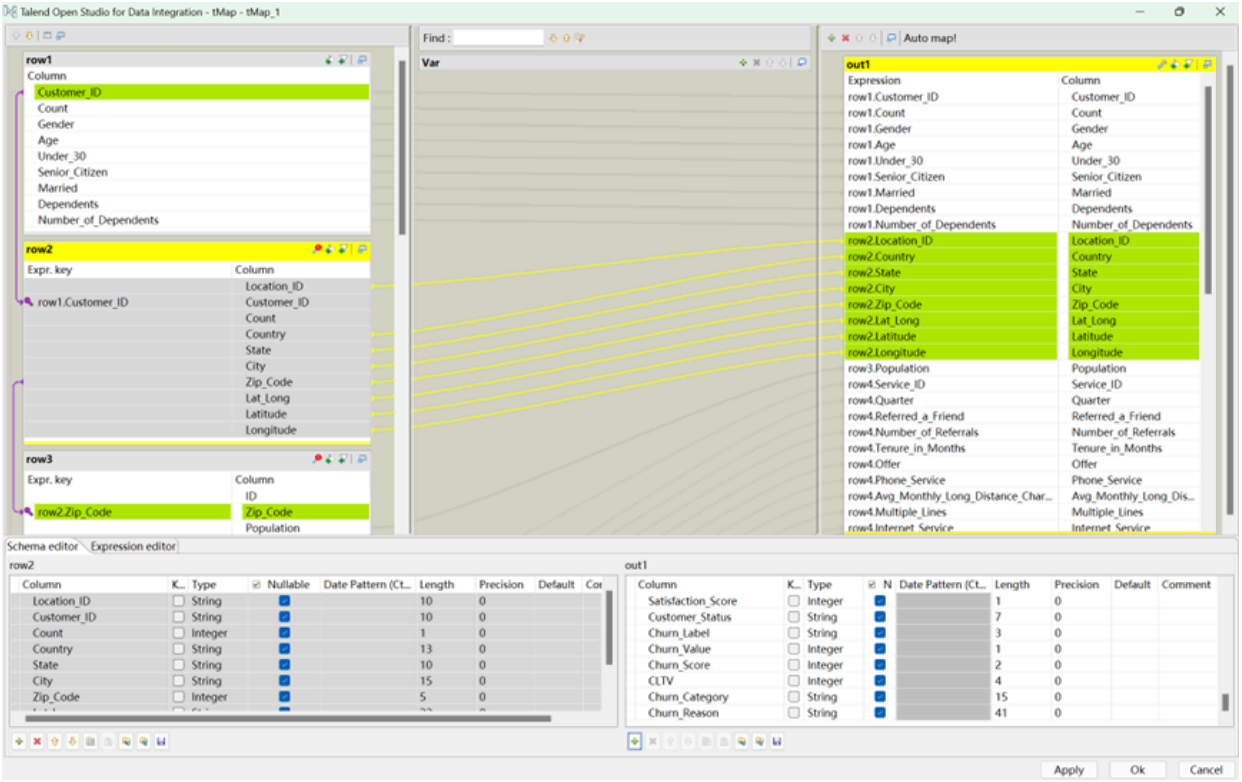

New Excel File
File - Step 4 of 4
Add a Schema on repository
Define the Schema
Name status
Comment
Schema
Click to update schema preview
Guess
Description of the Schema
Column K... Type N Date Pattern... Length Precisi... Defa... Comm...
Status_ID String 10 0
Customer_ID String 10 0
Count Integer 1 0
Quarter String 2 0
Satisfaction_Score Integer 1 0
< Back
Next >
Finish
Cancel

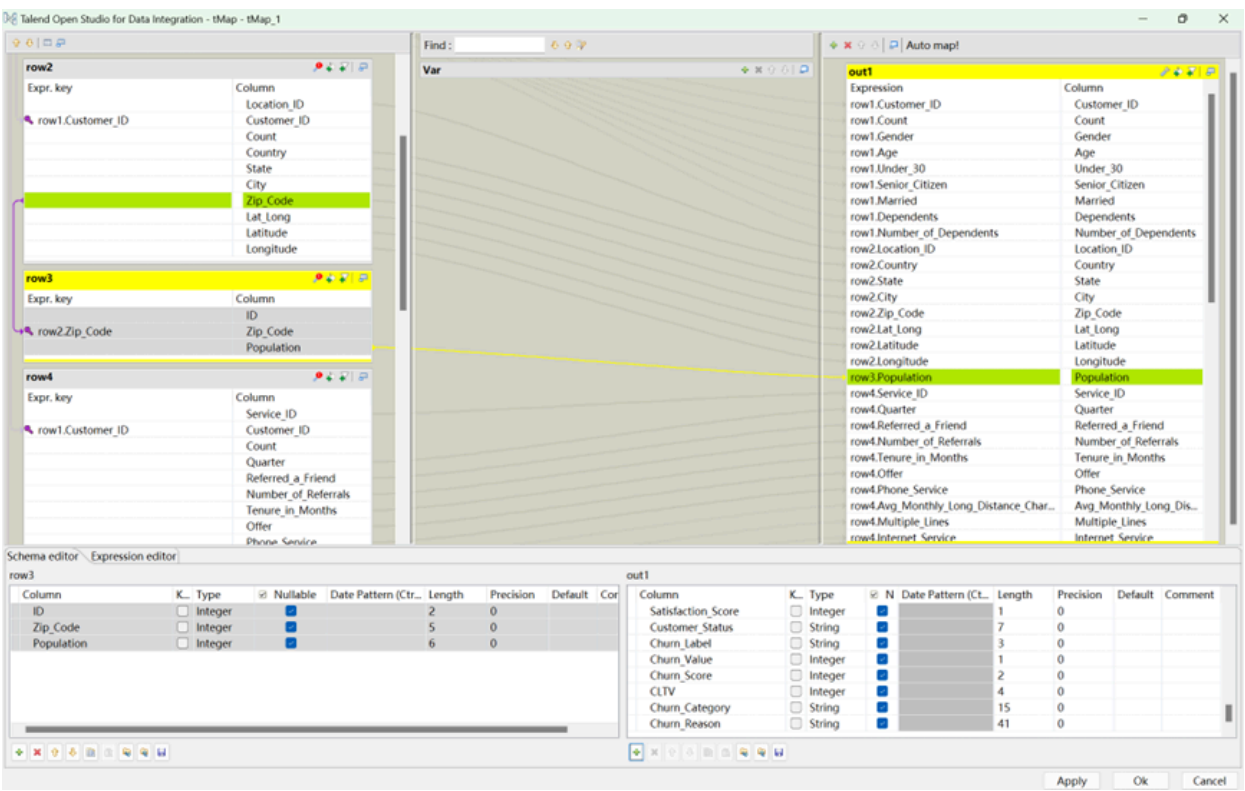
5. Add a tMap component and link all tFileInputExcel components to the tMap component:

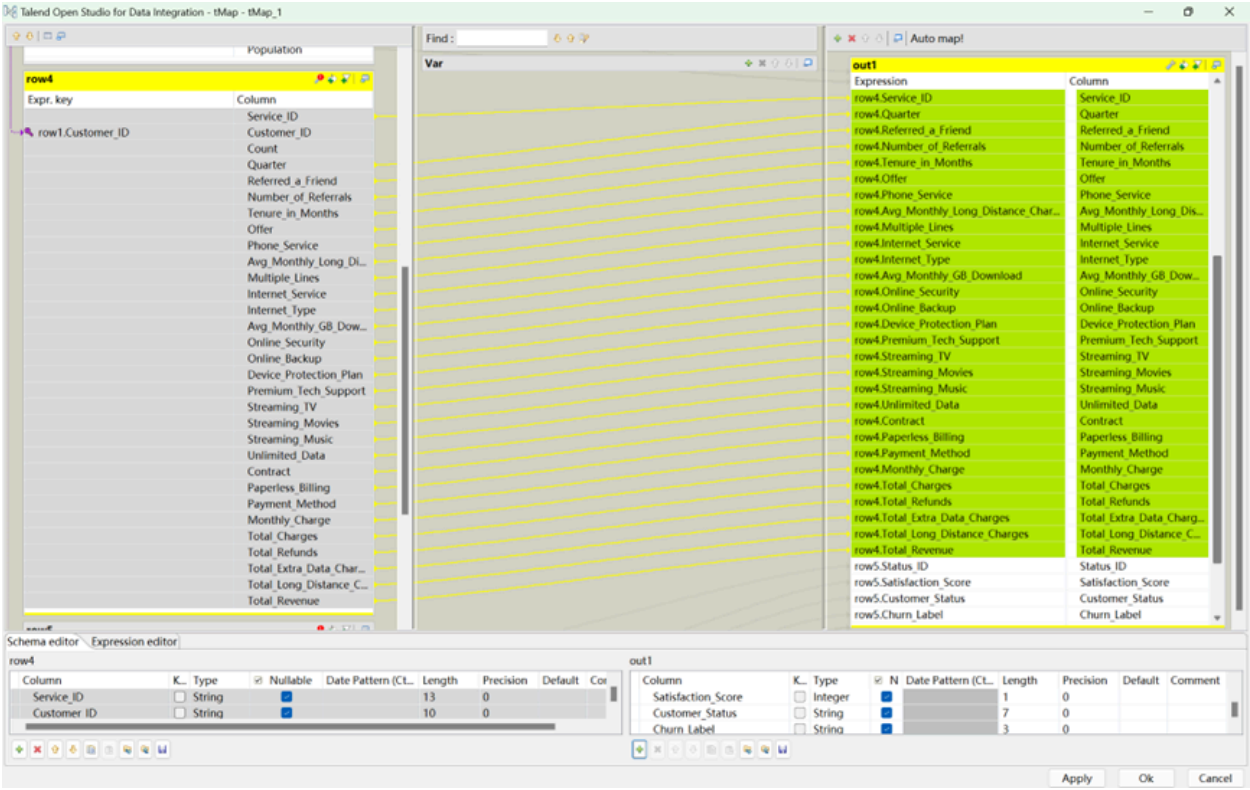


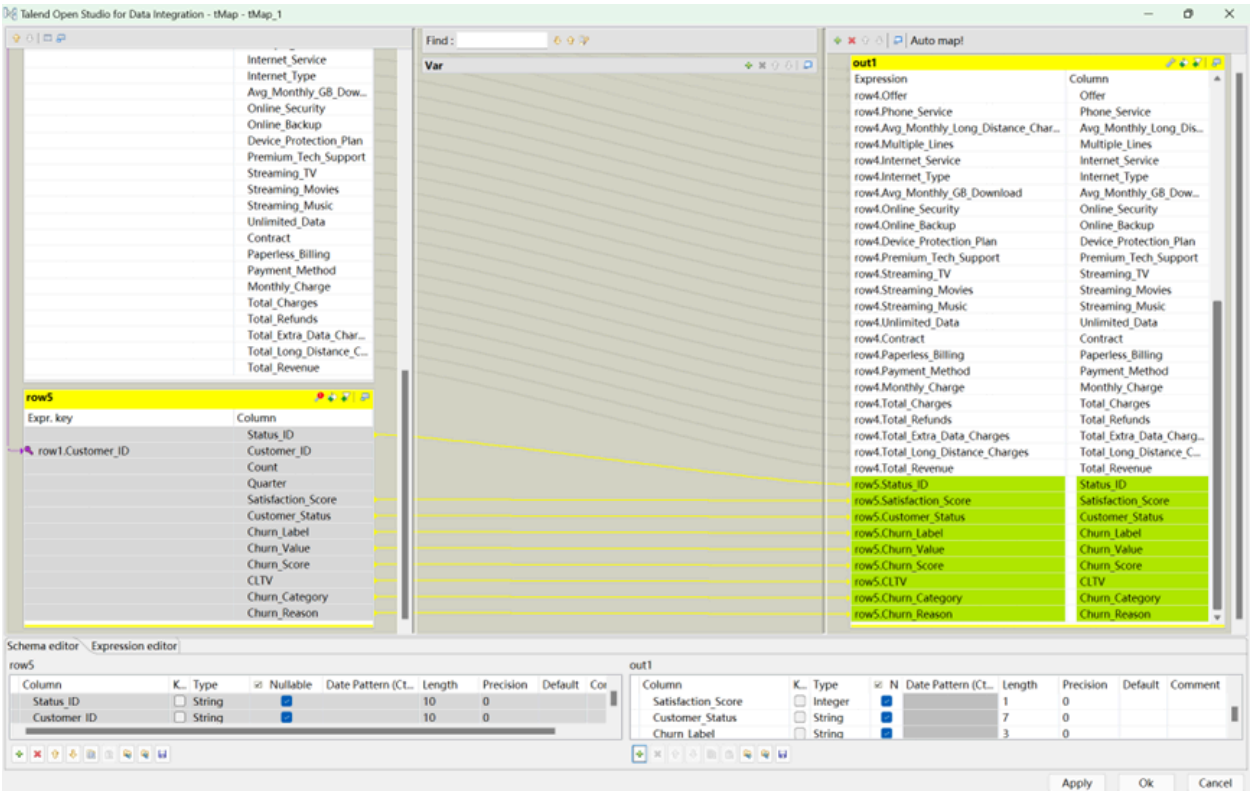
6. Joining all the data in the tMap editor:



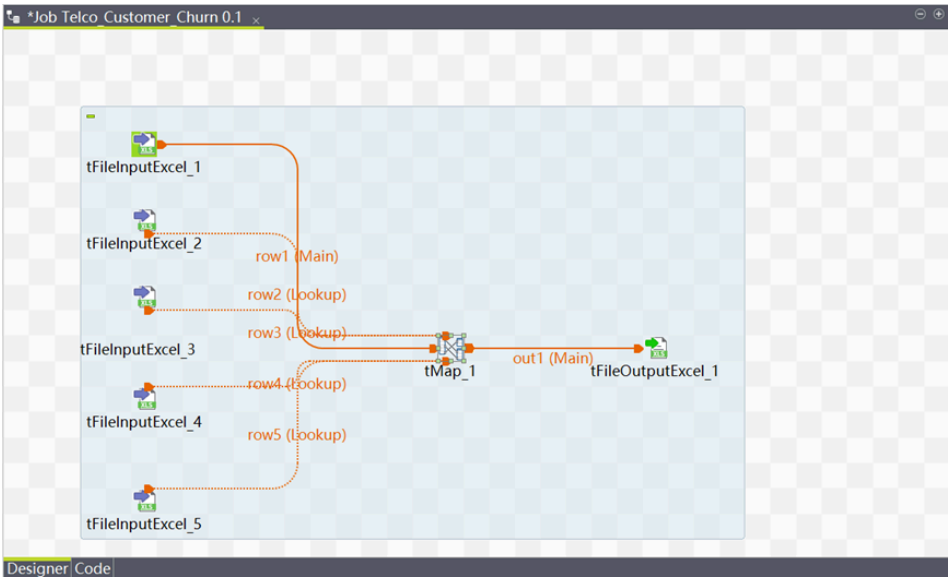

Talend Open Studio for Data Integration - tMap - tMap_1
row1
Column
Customer_ID
Count
Gender
Age
Under_30
Senior_Citizen
Married
Dependents
Number_of_Dependents
row2
Expr. key
Column
row1.Customer_ID
Location_ID
Customer_ID
Count
Country
State
City
Zip_Code
Lat_Long
Latitude
Longitude
row3
Expr. key
Column
row2.Zip_Code
ID
Zip_Code
Population
Var
out1
Expression
Column
row1.Customer_ID
row1.Count
row1.Gender
row1.Age
row1.Under_30
row1.Senior_Citizen
row1.Married
row1.Dependents
row1.Number_of_Dependents
row2.Location_ID
row2.Country
row2.State
row2.City
row2.Zip_Code
row2.Lat_Long
row2.Latitude
row2.Longitude
row3.Population
row4.Service_ID
row4.Quarter
row4.Referred_a_Friend
row4.Number_of_Referrals
row4.Tenure_in_Months
row4.Offer
row4.Phone_Service
row4.Avg_Monthly_Long_Distance_Char...
row4.Multiple_Lines
row4.Internet_Service
Schema editor
Expression editor
row2
Column
K...
Type
Nullable
Date Pattern (Ct...
Length
Precision
Default
Cor
Location_ID String 10 0
Customer_ID String 10 0
Count Integer 1 0
Country String 13 0
State String 10 0
City String 15 0
Zip_Code Integer 5 0
out1
Column
K...
Type
N
Date Pattern (Ct...
Length
Precision
Default
Comment
Satisfaction_Score Integer 1 0
Customer_Status String 7 0
Churn_Label String 3 0
Churn_Value Integer 1 0
Churn_Score Integer 2 0
CLTV Integer 4 0
Churn_Category String 15 0
Churn_Reason String 41 0
Apply
Ok
Cancel


Talend Open Studio for Data Integration - tMap - tMap_1
row2
Expr. key
Column
row1.Customer_ID
Location_ID
Customer_ID
Count
Country
State
City
Zip_Code
Lat_Long
Latitude
Longitude
row3
Expr. key
Column
row2.Zip_Code
ID
Zip_Code
Population
row4
Expr. key
Column
row1.Customer_ID
Service_ID
Customer_ID
Count
Quarter
Referred_a_Friend
Number_of_Referrals
Tenure_in_Months
Offer
Phone_Service
Var
out1
Expression
Column
row1.Customer_ID
row1.Count
row1.Gender
row1.Age
row1.Under_30
row1.Senior_Citizen
row1.Married
row1.Dependents
row1.Number_of_Dependents
row2.Location_ID
row2.Country
row2.State
row2.City
row2.Zip_Code
row2.Lat_Long
row2.Latitude
row2.Longitude
row3.Population
row4.Service_ID
row4.Quarter
row4.Referred_a_Friend
row4.Number_of_Referrals
row4.Tenure_in_Months
row4.Offer
row4.Phone_Service
row4.Avg_Monthly_Long_Distance_Char...
row4.Multiple_Lines
row4.Internet_Service
Schema editor
Expression editor
row3
Column
K...
Type
Nullable
Date Pattern (Ctr...
Length
Precision
Default
Cor
ID Integer 2 0
Zip_Code Integer 5 0
Population Integer 6 0
out1
Column
K...
Type
N
Date Pattern (Ct...
Length
Precision
Default
Comment
Satisfaction_Score Integer 1 0
Customer_Status String 7 0
Churn_Label String 3 0
Churn_Value Integer 1 0
Churn_Score Integer 2 0
CLTV Integer 4 0
Churn_Category String 15 0
Churn_Reason String 41 0
Apply
Ok
Cancel

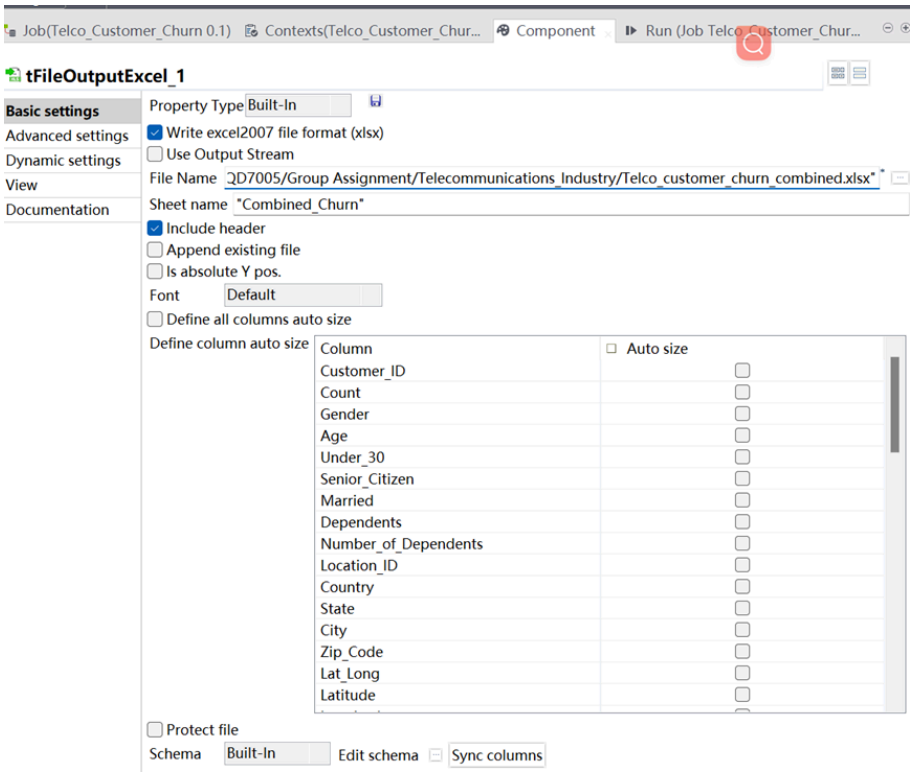

Talend Open Studio for Data Integration - tMap - tMap_1
Find :
Var
Auto map!
row4
Expr. key
Column
row1.Customer_ID
Service_ID
Customer_ID
Count
Quarter
Referred_a_Friend
Number_of_Referrals
Tenure_in_Months
Offer
Phone_Service
Avg_Monthly_Long_Di...
Multiple_Lines
Internet_Service
Internet_Type
Avg_Monthly_GB_Dow...
Online_Security
Online_Backup
Device_Protection_Plan
Premium_Tech_Support
Streaming_TV
Streaming_Movies
Streaming_Music
Unlimited_Data
Contract
Paperless_Billing
Payment_Method
Monthly_Charge
Total_Charges
Total_Refunds
Total_Extra_Data_Char...
Total_Long_Distance_C...
Total_Revenue
out1
Expression
Column
row4.Service_ID
row4.Quarter
row4.Referred_a_Friend
row4.Number_of_Referrals
row4.Tenure_in_Months
row4.Offer
row4.Phone_Service
row4.Avg_Monthly_Long_Distance_Char...
row4.Multiple_Lines
row4.Internet_Service
row4.Internet_Type
row4.Avg_Monthly_GB_Download
row4.Online_Security
row4.Online_Backup
row4.Device_Protection_Plan
row4.Premium_Tech_Support
row4.Streaming_TV
row4.Streaming_Movies
row4.Streaming_Music
row4.Unlimited_Data
row4.Contract
row4.Paperless_Billing
row4.Payment_Method
row4.Monthly_Charge
row4.Total_Charges
row4.Total_Refunds
row4.Total_Extra_Data_Charges
row4.Total_Long_Distance_Charges
row4.Total_Revenue
row5.Status_ID
row5.Satisfaction_Score
row5.Customer_Status
row5.Churn_Label
Status_ID
Satisfaction_Score
Customer_Status
Churn_Label
Schema editor
Expression editor
row4
Column
K...
Type
Nullable
Date Pattern (Ct...
Length
Precision
Default
Service_ID
String
13
0
Customer_ID
String
10
0
out1
Column
K...
Type
N
Date Pattern (Ct...
Length
Precision
Default
Comment
Satisfaction_Score
Integer
1
0
Customer_Status
String
7
0
Churn_Label
String
3
0
Apply
Ok
Cancel


Talend Open Studio for Data Integration - tMap - tMap_1
Find :
Var
Auto map!
Internet_Service
Internet_Type
Avg_Monthly_GB_Dow...
Online_Security
Online_Backup
Device_Protection_Plan
Premium_Tech_Support
Streaming_TV
Streaming_Movies
Streaming_Music
Unlimited_Data
Contract
Paperless_Billing
Payment_Method
Monthly_Charge
Total_Charges
Total_Refunds
Total_Extra_Data_Char...
Total_Long_Distance_C...
Total_Revenue
row5
Expr. key
Column
row1.Customer_ID
Status_ID
Customer_ID
Count
Quarter
Satisfaction_Score
Customer_Status
Churn_Label
Churn_Value
Churn_Score
CLTV
Churn_Category
Churn_Reason
out1
Expression
Column
row4.Offer
row4.Phone_Service
row4.Avg_Monthly_Long_Distance_Char...
row4.Multiple_Lines
row4.Internet_Service
row4.Internet_Type
row4.Avg_Monthly_GB_Download
row4.Online_Security
row4.Online_Backup
row4.Device_Protection_Plan
row4.Premium_Tech_Support
row4.Streaming_TV
row4.Streaming_Movies
row4.Streaming_Music
row4.Unlimited_Data
row4.Contract
row4.Paperless_Billing
row4.Payment_Method
row4.Monthly_Charge
row4.Total_Charges
row4.Total_Refunds
row4.Total_Extra_Data_Charges
row4.Total_Long_Distance_Charges
row4.Total_Revenue
row5.Status_ID
row5.Satisfaction_Score
row5.Customer_Status
row5.Churn_Label
row5.Churn_Value
row5.Churn_Score
row5.CLTV
row5.Churn_Category
row5.Churn_Reason
Offer
Phone_Service
Avg_Monthly_Long_Dis...
Multiple_Lines
Internet_Service
Internet_Type
Avg_Monthly_GB_Dow...
Online_Security
Online_Backup
Device_Protection_Plan
Premium_Tech_Support
Streaming_TV
Streaming_Movies
Streaming_Music
Unlimited_Data
Contract
Paperless_Billing
Payment_Method
Monthly_Charge
Total_Charges
Total_Refunds
Total_Extra_Data_Charg...
Total_Long_Distance_C...
Total_Revenue
Status_ID
Satisfaction_Score
Customer_Status
Churn_Label
Churn_Value
Churn_Score
CLTV
Churn_Category
Churn_Reason
Schema editor
Expression editor
row5
Column
K...
Type
Nullable
Date Pattern (Ct...
Length
Precision
Default
Status_ID
String
10
0
Customer_ID
String
10
0
out1
Column
K...
Type
N
Date Pattern (Ct...
Length
Precision
Default
Comment
Satisfaction_Score
Integer
1
0
Customer_Status
String
7
0
Churn_Label
String
3
0
Apply
Ok
Cancel

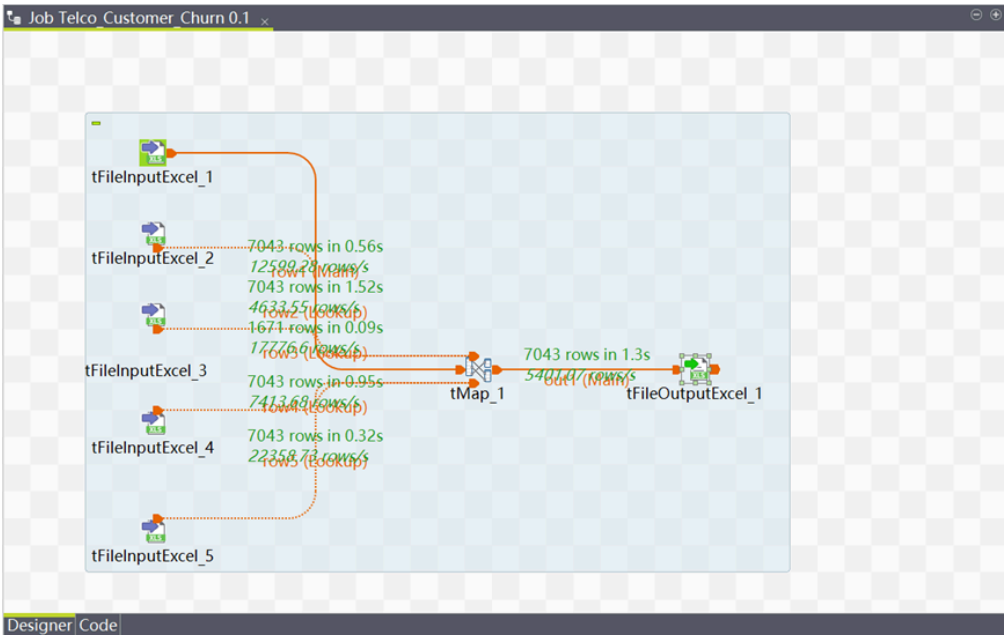
7. Add a tFileOutputExcel component and link the tMap component to the tFileOutputExcel component:



8. Configure the output component to write the results to a new Excel file, Telco_customer_churn_combined.xlsx.



9. Run and execute the ETL job:



After using Talend studio, we successfully combined the five excel files into one new excel file, named Telco_customer_churn_combined.

4.3 Sample (SEMMA: Step 1)

No sampling was conducted at this stage as the combined dataset consists of 7043 rows and 33 columns which is manageable using tools such as Talend and SAS. This allows highly granular data to be maintained, enabling a better understanding of the dataset and insights to be drawn from the full dataset. Furthermore, the risk of sampling bias can be entirely eliminated. This approach is similar to the sample report "Group 5 Group Assignment report_v1" where the entire dataset was used in the Explore stage.

4.4 Explore (SEMMA: Step 2)

In the next step, Univariate, Bivariate, Trivariate and Multivariate Exploratory Data Analysis (EDA) was conducted using SAS Enterprise Miner to better understand the dataset.

4.4.1 Importing and Summary Statistics

The combined dataset was first imported into SAS Enterprise Miner using the File Import node.

Figure 4.4.1: SAS Diagram

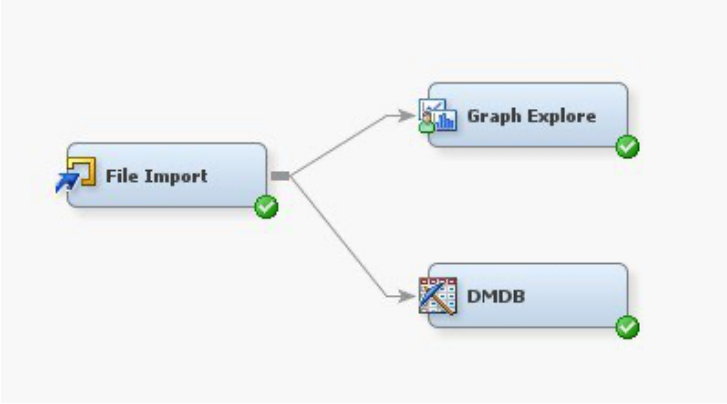


Figure 4.4.2: File Import Settings

Property	Value
General	
Node ID	FIMPORT
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Import File	C:\Users\zyang\Downloa
Maximum Rows to Import	1000000
Maximum Columns to Impor	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500
File Location	Local
File Type	xlsx
Advanced Advisor	No
Rerun	No
Score	
Role	Train

Figure 4.4.5: Variable Role Classification

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
CLTV	Input	Interval	No		No	.	.
Churn_Label	Target	Nominal	No		No	.	.
Churn_Reason	Input	Nominal	No		No	.	.
Churn_Score	Rejected	Interval	No		Yes	.	.
Churn_Value	Rejected	Binary	No		Yes	.	.
City	Input	Nominal	No		No	.	.
Contract	Input	Nominal	No		No	.	.
Count	Rejected	Interval	No		Yes	.	.
Country	Input	Nominal	No		No	.	.
CustomerID	Rejected	Nominal	No		Yes	.	.
Dependents	Input	Nominal	No		No	.	.
Device_Protectic	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
Internet_Service	Input	Nominal	No		No	.	.
Lat_Long	Rejected	Nominal	No		Yes	.	.
Latitude	Rejected	Interval	No		Yes	.	.
Longitude	Rejected	Interval	No		Yes	.	.
Monthly_Charge	Input	Interval	No		No	.	.
Multiple_Lines	Input	Nominal	No		No	.	.
Online_Backup	Input	Nominal	No		No	.	.
Online_Security	Input	Nominal	No		No	.	.
Paperless_Billing	Input	Nominal	No		No	.	.
Partner	Input	Nominal	No		No	.	.
Payment_Metho	Input	Nominal	No		No	.	.
Phone_Service	Input	Nominal	No		No	.	.
Senior_Citizen	Input	Nominal	No		No	.	.
State	Input	Nominal	No		No	.	.
Streaming_Movie	Input	Nominal	No		No	.	.
Streaming_TV	Input	Nominal	No		No	.	.
Tech_Support	Input	Nominal	No		No	.	.
Tenure_Months	Input	Interval	No		No	.	.
Total_Charges	Input	Interval	No		No	.	.
Zip_Code	Input	Nominal	No		No	.	.

After import, SAS automatically assigned roles for all variables. These automatic role classifications were adjusted to better represent the dataset and achieve the data mining goals. First, Churn_Label was set as the Target. The variables Churn_Score, Churn_Value, Count, CustomerID, Lat_Long, Latitude and Longitude were rejected because Churn_Score is a predicted value while the other variables were not deemed to be useful to the analysis based on the variable descriptions. The dataset now contains 7043 rows and 27 columns.

Summary statistics for all variables were then produced using the DMDB node as shown in Figure 4.4.3 and Figure 4.4.4 below.

Figure 4.4.3: Summary Statistics of Class Variables

Class Variables

Variable	Label	Type	Number of Levels	Missing
Churn_Reason	Churn Reason	C	20	5174
Churn_Value	Churn Value	N	2	0
City	City	C	26	0
Contract	Contract	C	3	0
Country	Country	C	1	0
Dependents	Dependents	C	2	0
Device_Protection	Device Protection	C	3	0
Gender	Gender	C	2	0
Internet_Service	Internet Service	C	3	0
Multiple_Lines	Multiple Lines	C	3	0
Online_Backup	Online Backup	C	3	0
Online_Security	Online Security	C	3	0
Paperless_Billing	Paperless Billing	C	2	0
Partner	Partner	C	2	0
Payment_Method	Payment Method	C	4	0
Phone_Service	Phone Service	C	2	0
Senior_Citizen	Senior Citizen	C	2	0
State	State	C	1	0
Streaming_Movies	Streaming Movies	C	3	0
Streaming_TV	Streaming TV	C	3	0
Tech_Support	Tech Support	C	3	0
Zip_Code	Zip Code	N	26	0

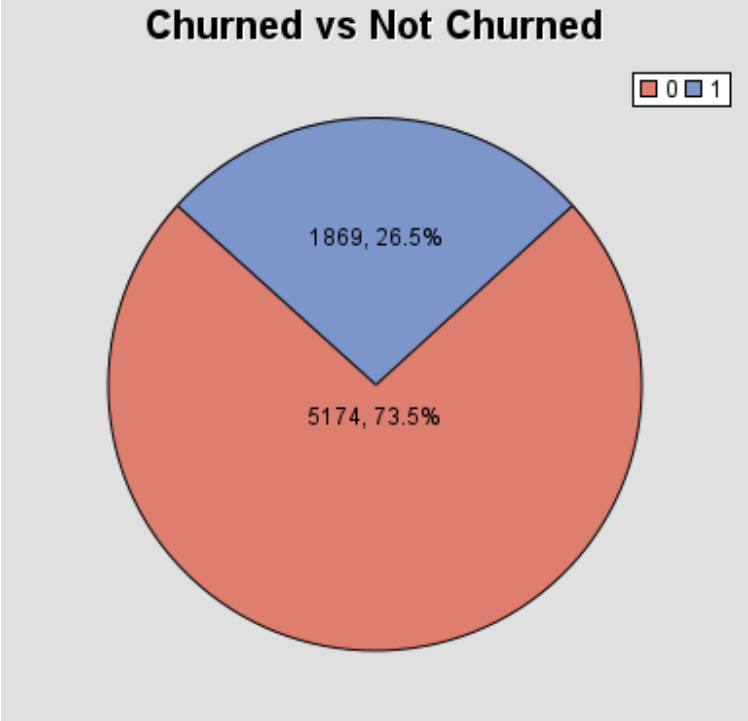
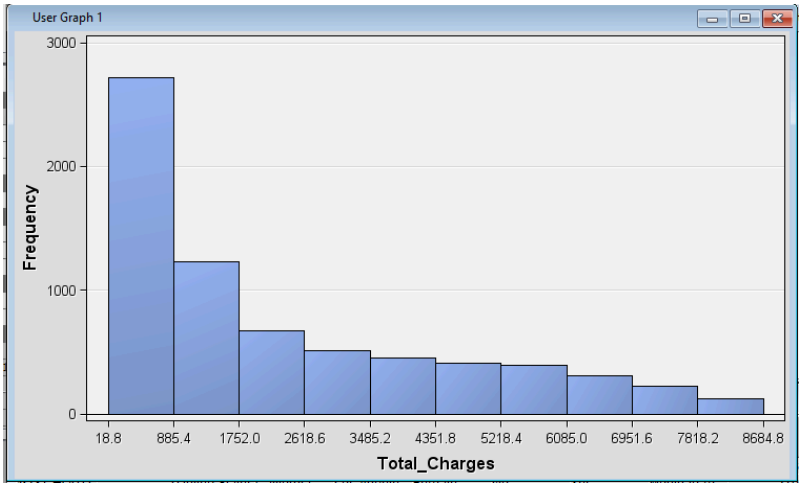
Figure 4.4.4: Summary Statistics of Interval Variables

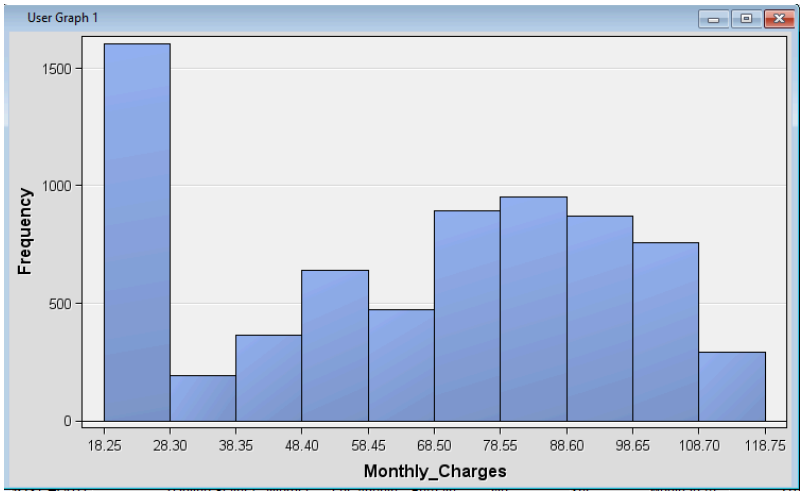
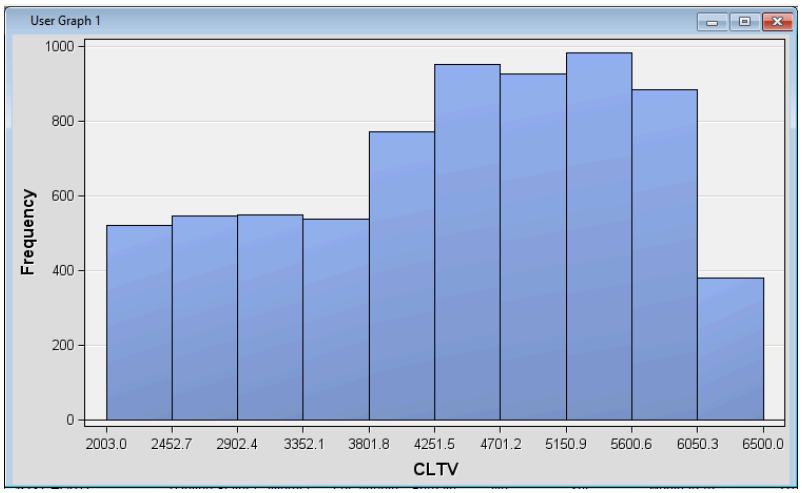
Interval Variables

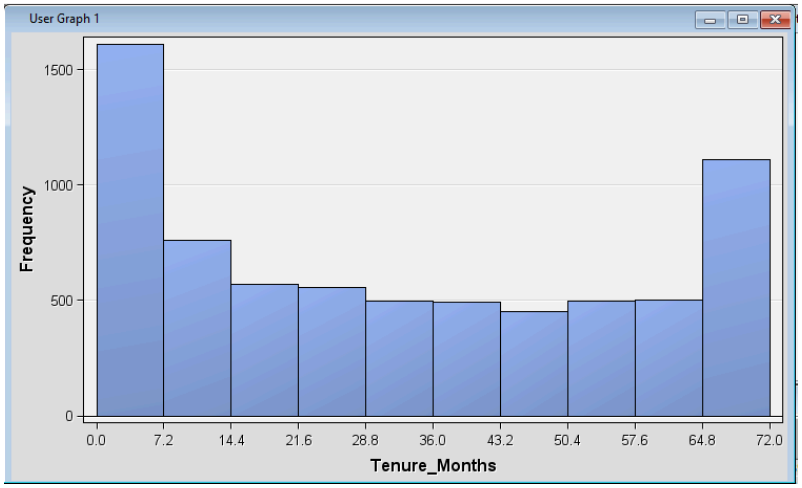
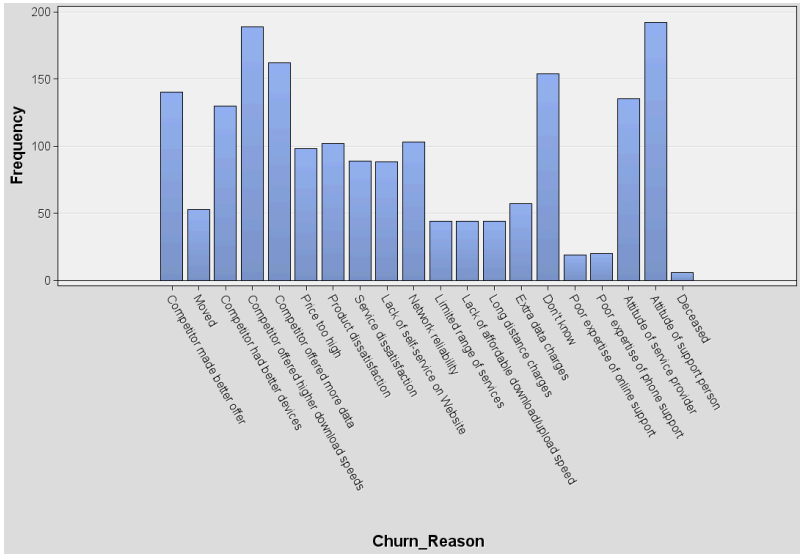
Variable	Label	Missing	N	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
CLTV	CLTV	0	7043	2003	6500	4400.296	1183.057	-0.3116	-0.93403
Monthly_Ch...	Monthly Ch...	0	7043	18.25	118.75	64.76169	30.09005	-0.22052	-1.25726
Tenure_Mo...	Tenure Mon...	0	7043	0	72	32.37115	24.55948	0.23954	-1.38737
Total_Char...	Total Charg...	0	7043	18.8	8684.8	2279.799	2266.73	0.963316	-0.22849

The only variable with missing values is Churn_Reason, with 5174 missing values which represents ~73% of the data. Since the percentage of missing values is high, this variable will only be investigated in the Explore phase and will be dropped before the Model phase.

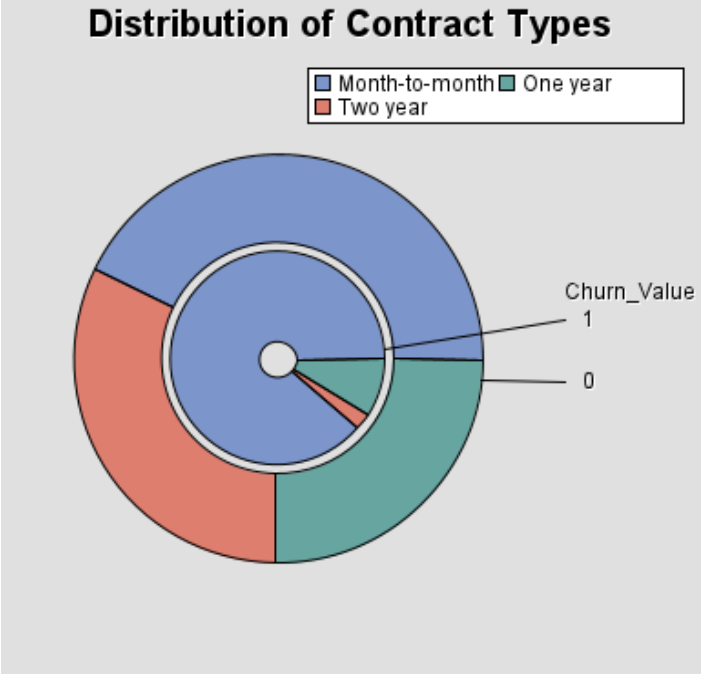
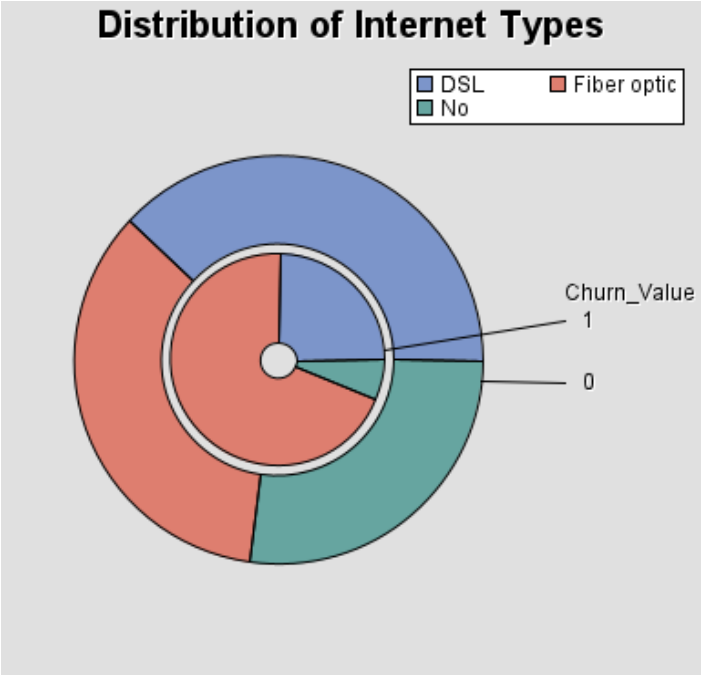
4.4.2 Univariate Analysis

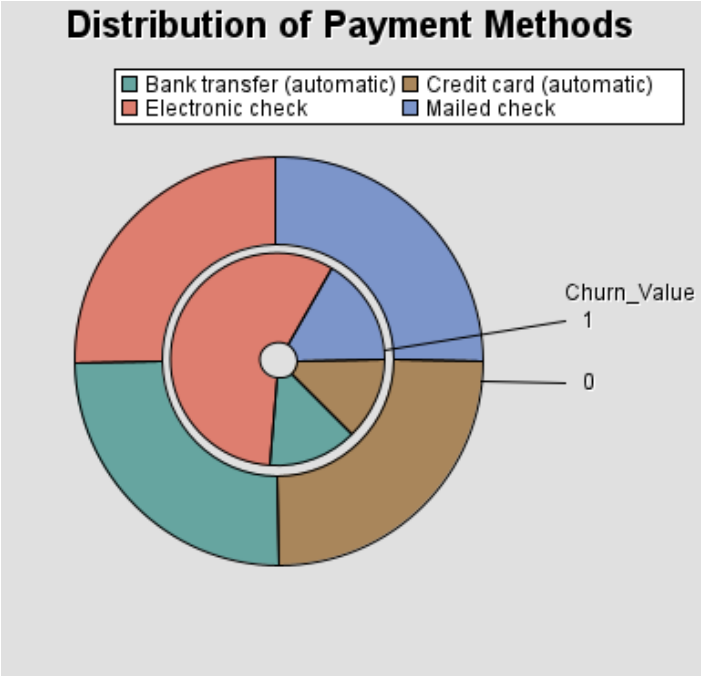
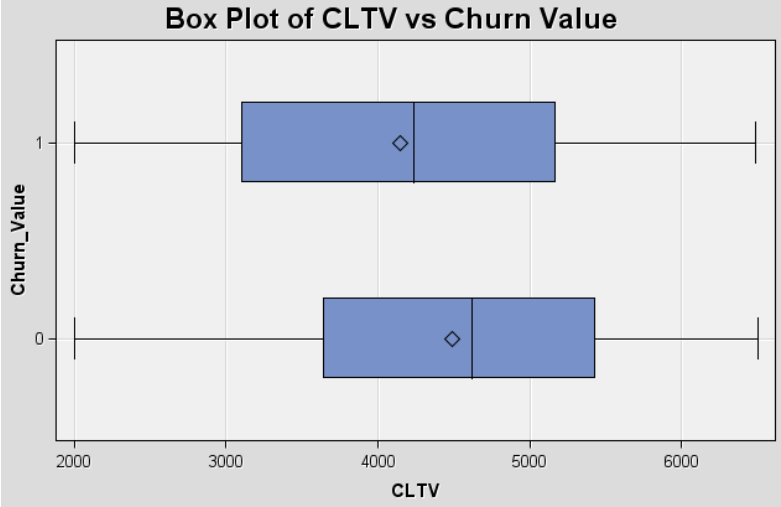
Variable	Chart	Highlights
Churn Label	Figure 4.4.4: Distribution of customers who churned. 	- Target variable. - Dataset is imbalanced with positive churned cases representing only 26.5%.
Total Charges	Figure 4.4.5: Distribution of Total Charges. 	- Right skewed distribution. - Most customers have Total Charges below 885.4.

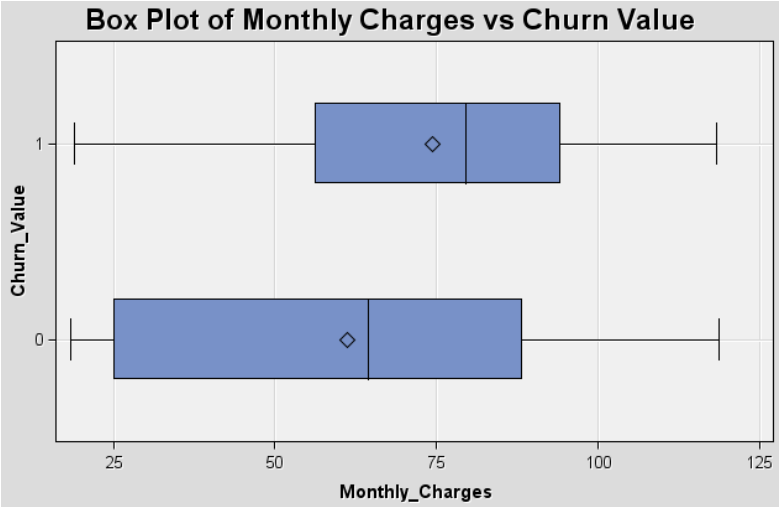
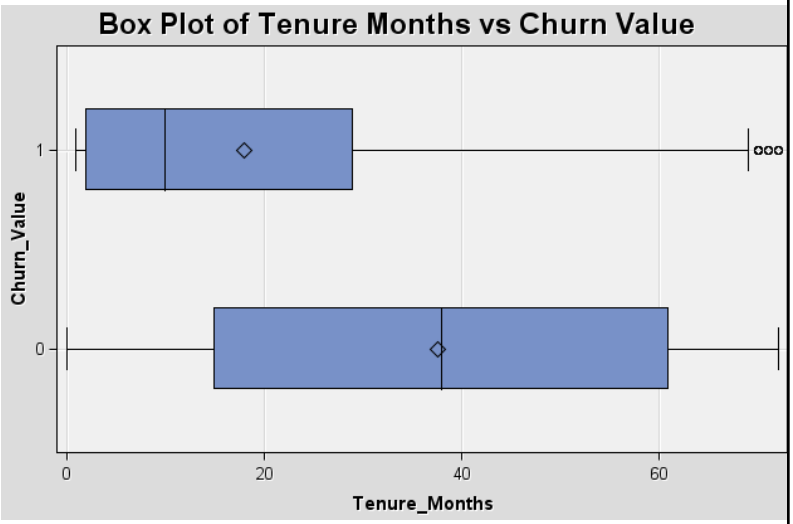
Monthly Charges	Figure 4.4.6: Distribution of Monthly Charges.  	- A peak is seen for Monthly Charges between 18.25 and 28.30.
CLTV	Figure 4.4.7: Distribution of CLTV.  	- Left skewed distribution. - Most customers have a higher CLTV value.

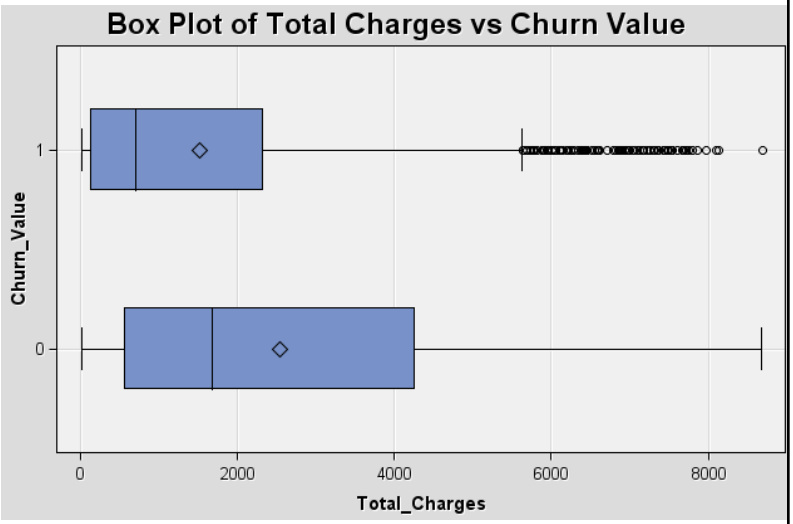
Tenure Months	Figure 4.4.8: Distribution of Tenure in months.  	- Two noticeable peaks are present. - Many customers either have a very short tenure or a very long tenure.
Churn Reason	Figure 4.4.9: Distribution of Churn Reasons.  	- Customers mostly churn because they received unsatisfactory support or were offered better prices, speeds or quota by other competitors.

4.4.3 Bivariate Analysis

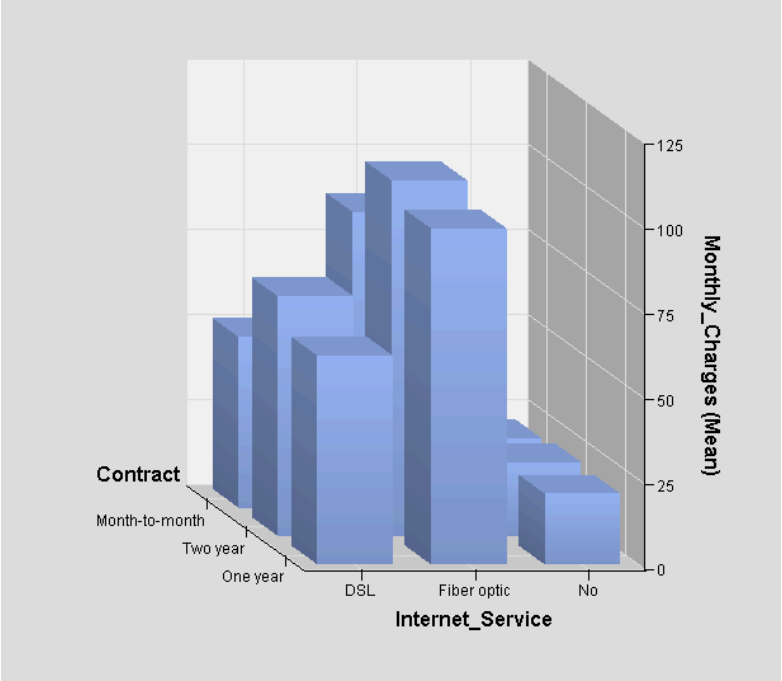
Variable	Chart	Findings
Contract, Churn	Figure 4.4.10: Pie Chart of Contract Types vs Churn. 	- A higher percentage of customers who churn have month-to-month contract types.
Internet Type, Churn	4.4.11: Pie Chart of Internet Types vs Churn. 	- A higher percentage of customers who churn have fiber optic internet.

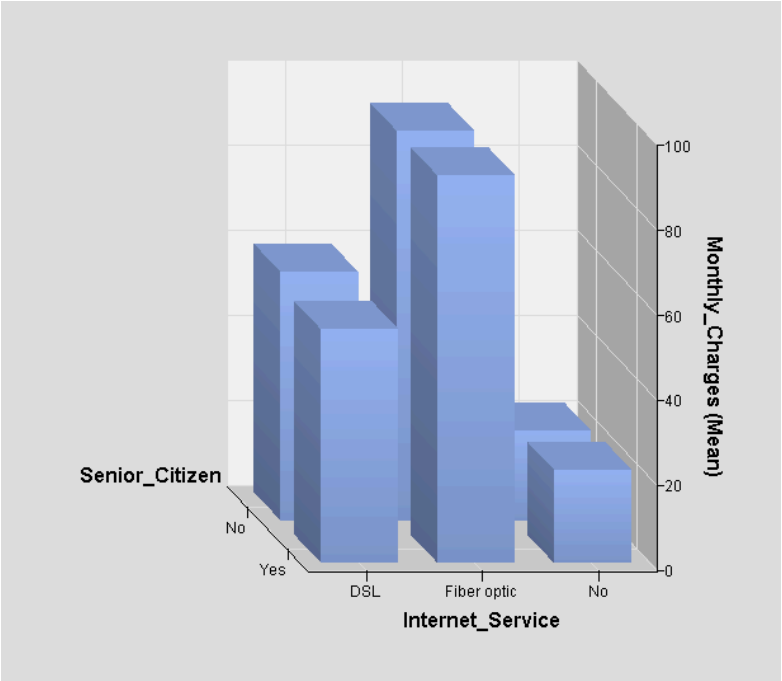
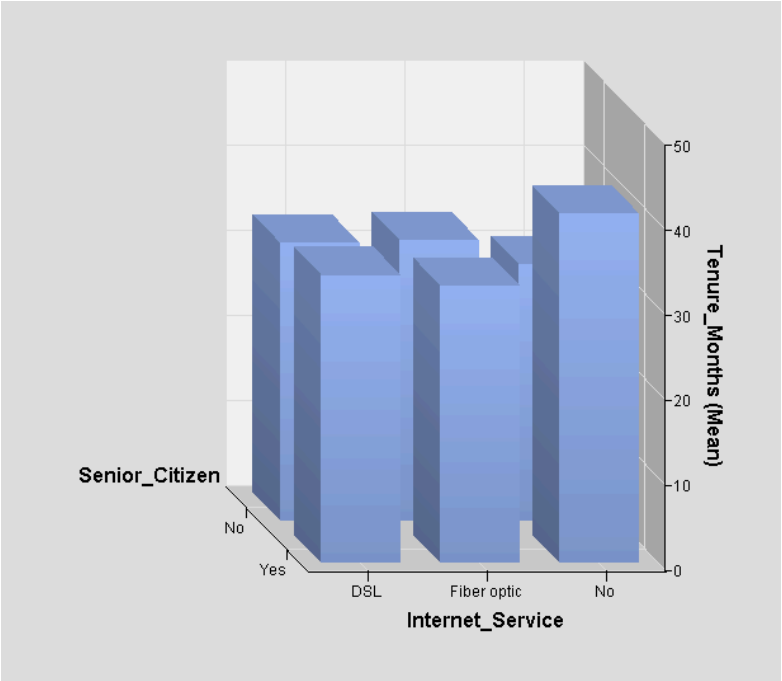
Payment Methods, Churn	Figure 4.4.12: Pie Chart of Payment Methods vs Churn.  	- A larger percentage of customers who churn pay by electronic check.
CLTV, Churn	Figure 4.4.13: Box Plots of CLTV vs Churn.  	- CLTV of customers who churn have a lower interquartile range (IQR).

Monthly Charges, Churn	Figure 4.4.13: Box Plots of Monthly Charges vs Churn.  	- Monthly charges borne by customers who churn have a higher IQR than those that do not churn. - Customers who churn mostly pay higher monthly charges.
Tenure Months, Churn	Figure 4.4.14: Box Plots of Tenure vs Churn  	- Tenure of customers who churn have a lower IQR than those that do not churn. - Customers who churn mostly have a lower tenure.

Total Charges, Churn	Figure 4.4.15: Box Plots of Total Charges vs Churn.  	- Total Charges of customers who churn have a lower IQR than those that do not churn. However, there are some outliers with very high CLTV values.

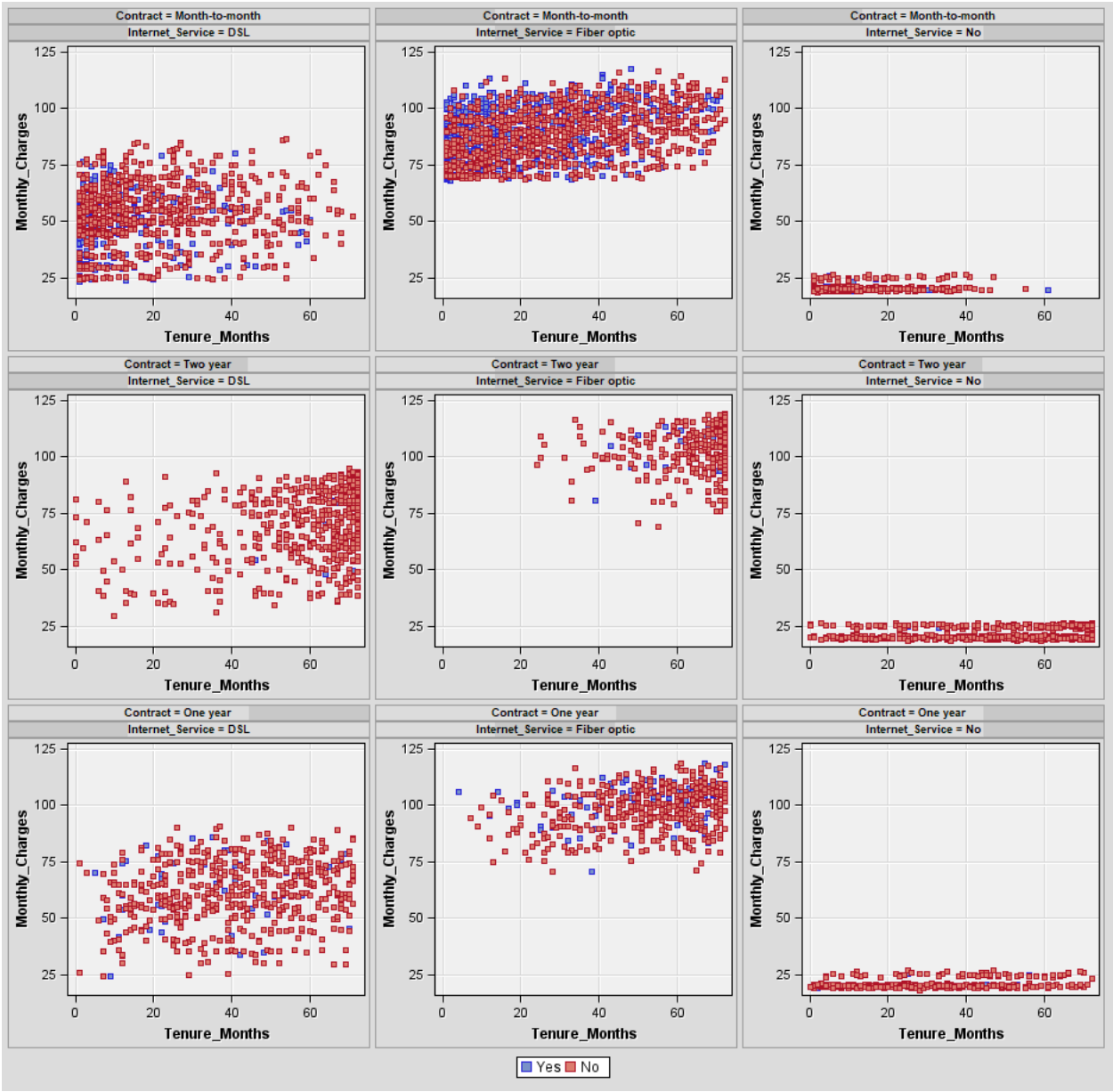
4.4.4 Trivariate Analysis

Variable	Chart	Findings
Contract, Internet Service, Monthly Charges	Figure 4.4.16: 3D Bar Chart of Contract type vs Internet Service vs Tenure 	- Fiber optic internet service commands the highest monthly charges. - Two year contracts command a higher average price than month-to-month or one year contracts for both DSL and fiber optic internet types.

Internet Service, Senior Citizen, Monthly Charges	Figure 4.4.17: 3D Bar Chart of Internet Service vs Senior Citizen vs Monthly Charges  	- Senior citizens pay slightly less on average for DSL internet.
Internet Service, Senior Citizen, Tenure	Figure 4.4.18 3D Chart of Internet Service vs Senior Citizen vs Tenure  	- Senior citizens that do not subscribe to any internet service have the longest average tenure.

4.4.5 Multivariate Analysis

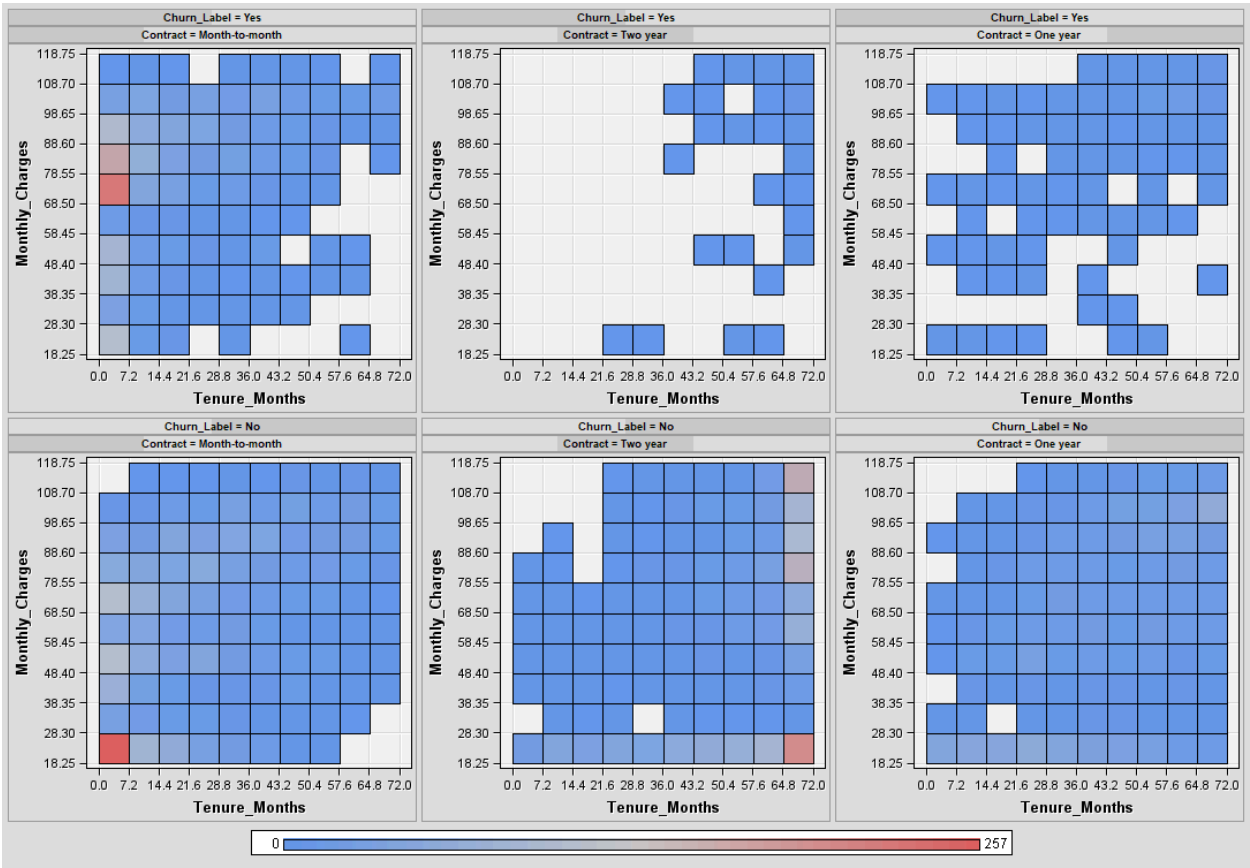
Figure 4.4.19: Lattice Plot of Tenure vs Monthly Charges for All Contract Types and Internet Service Types.



Scatter plots of Monthly Charges vs Tenure were plotted for each combination of Contract type and Internet Service. Red dots represent customers that did not churn while blue dots represent customers that did. From the plots, it can be seen that monthly charges are low for all customers without internet service and remain fairly constant even as

tenure increases. On the other hand, monthly charges for DSL and fiber optic internet tend to increase slightly as tenure increases. From the aspect of churn, no obvious trend could be seen.

Figure 4.4.20: Heatmaps of Tenure vs Monthly Charges for All Contract Types and Churn



Heatmaps of Tenure vs Monthly Charges were also plotted in a lattice to compare the relationship between the two variables for different Contract types and Churn Labels. One observation can be seen is that for month-to-month contract customers who churn, more of them are low tenure customers with relatively high monthly charges while month-to-month customers that did not churn mostly had low tenure and low monthly charges.

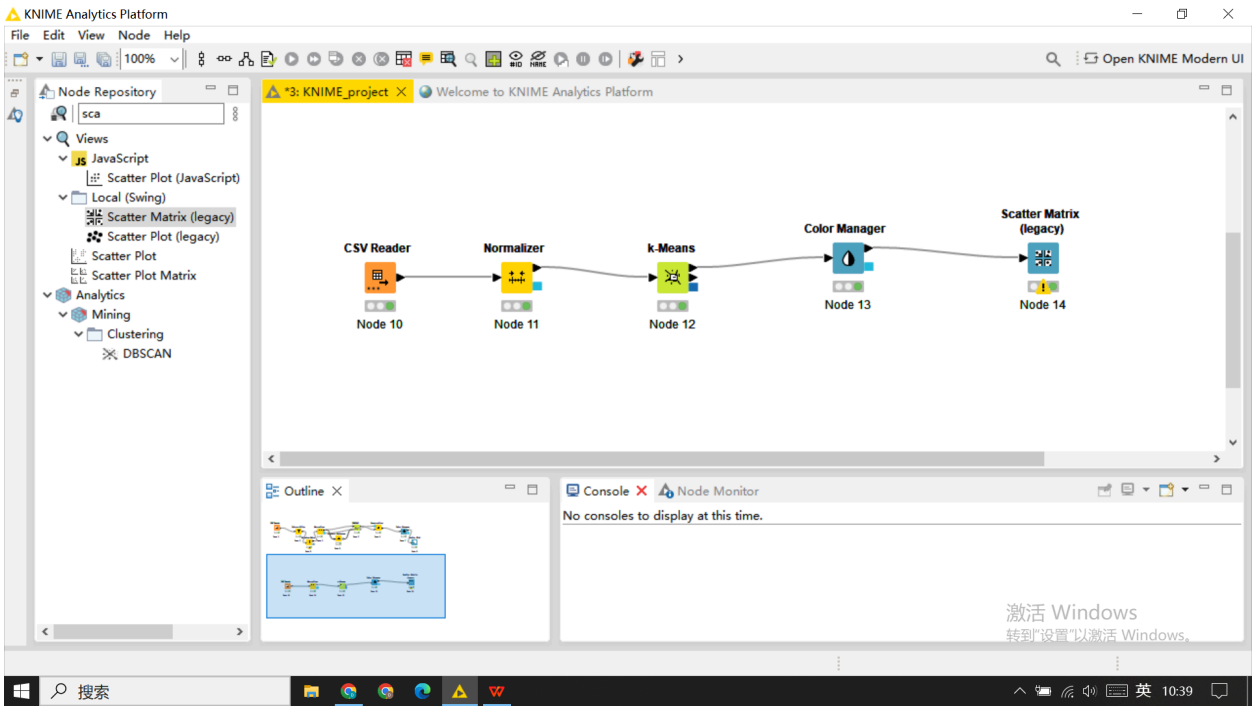
4.4.6 Clustering Analysis

We performed clustering analysis on customers using KNIME and SAS respectively, and the results are as follows.

a. Knime

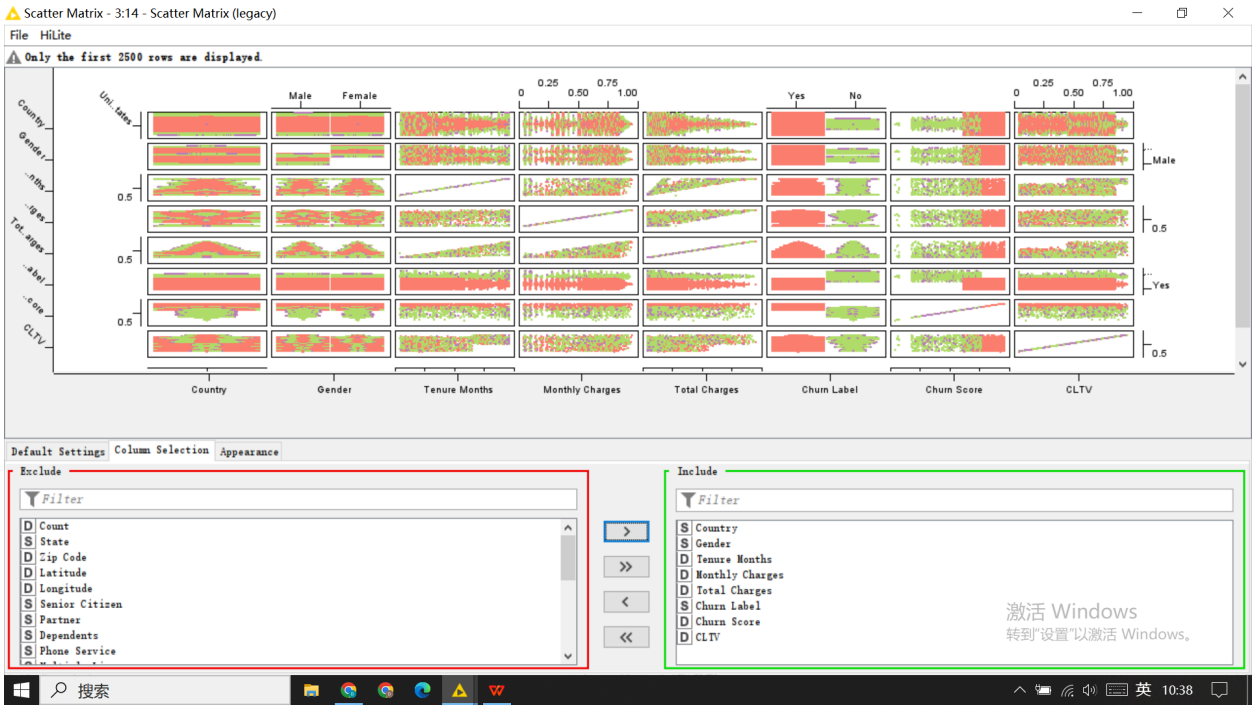
Figure 4.4.21 shows a simple workflow to perform customer segmentation using KNIME software applying the K-Means clustering process. Since there are many variables, we chose 'scatter matrix' in this project.

Figure 4.4.21: Workflow of clustering in KNIME



The results are as follows:

Figure 4.4.22: Output of clustering in KNIME

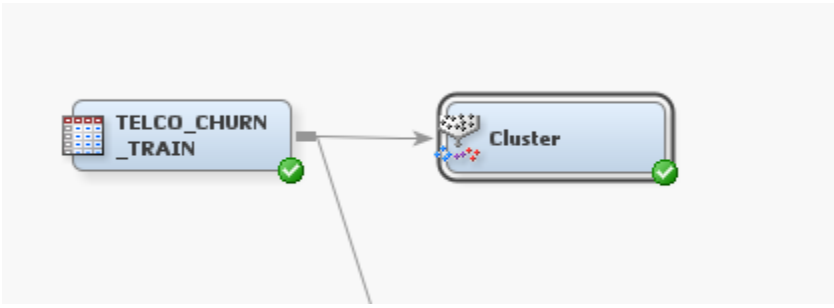


As shown in Figure 4.4.22, some variables like State, Latitude, and Longitude are excluded in the current scatter matrix, while Country, Gender, Tenure Months, Monthly Charges, Total Charges Churn Label, Churn Score, and CLTV are included, indicating that they are related to each others.

b. SAS

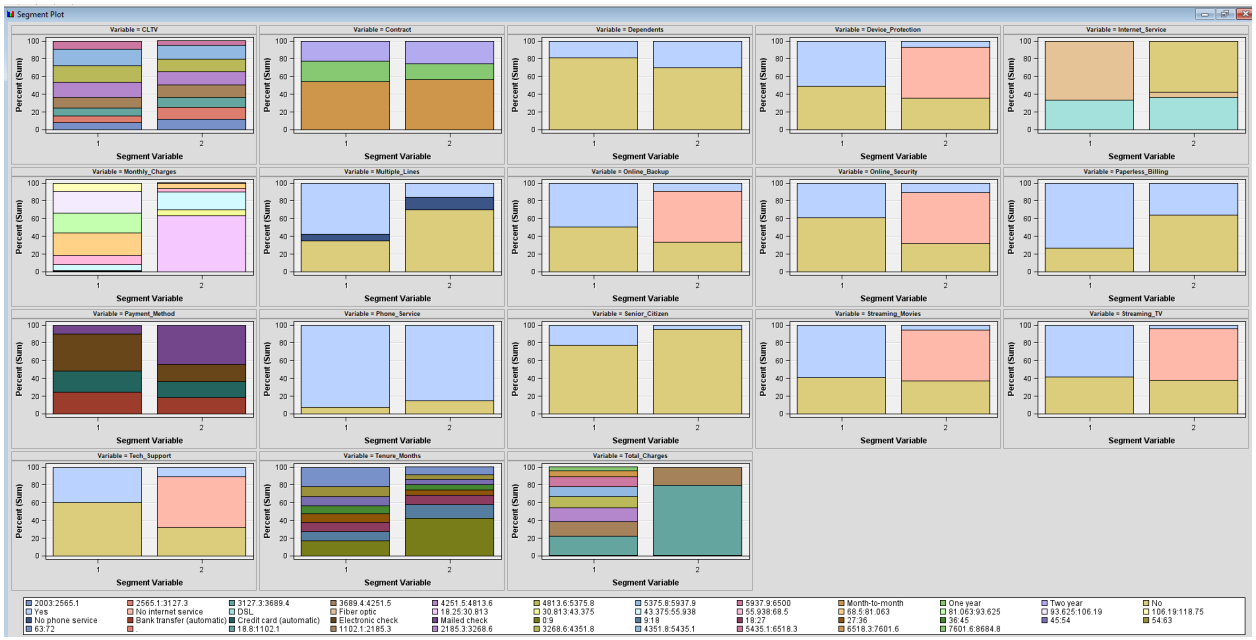
Figure 4.4.23 shows the workflow of clustering in SAS Enterprise Miner, it provides the output of variable importance to measure the ability of each variable to partition or split the dataset.

Figure 4.4.23: Workflow of clustering in SAS:



The result are as follows:

Figure 4.4.23&24: Output of clustering in SAS



Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
Monthly_Charges	Monthly Charges	5	10	1.00000
Internet_Service	Internet Service	0	4	0.88142
Online_Security	Online Security	0	4	0.87461
Streaming_Movies	Streaming Movies	1	5	0.87205
Online_Backup	Online Backup	0	3	0.87116
Device_Protection	Device Protection	0	3	0.87059
Total_Charges	Total Charges	7	6	0.49317
Tenure_Months	Tenure Months	0	12	0.31474
Contract		0	6	0.22059
Paperless_Billing	Paperless Billing	5	0	0.19592
Streaming_TV	Streaming TV	1	6	0.17169
Dependents		0	2	0.16625
CLTV		0	7	0.13703
Phone_Service	Phone Service	2	1	0.12002
Multiple_Lines	Multiple Lines	0	3	0.11937
Tech_Support	Tech Support	0	2	0.11599
Senior_Citizen	Senior Citizen	2	0	0.09589
Payment_Method	Payment Method	0	2	0.07262

As shown in Figure 4.4.24, “Importance” represents the relative importance of this variable in the clustering process. The closer the value is to 1, the more important the variable is in the partitioned dataset. Monthly_Charges is rated the most important variable (with an importance of 1.00000), indicating that it is the most critical split variable in the clustering model. At the bottom of the list, we can see that Payment_Method has the lowest importance (0.07262).

4.5 Modify (SEMMA: Step 3)

4.5.1 Fixing Misspelling and Data Errors

We imported the dataset in Talend Data Preparation and found a spelling error in the ‘Payment Method’ column as shown in Figure 4.5.1. For example 'Bank transfer (automatci)', 'Bank transfer (automatic)', 'Credit card (automaitc)', etc. We use the ‘FOUND AND GROUP SIMLIAR TEXT’ function to correct spelling errors.

Figure 4.5.1: Misspelling error in ‘Payment Method’ column

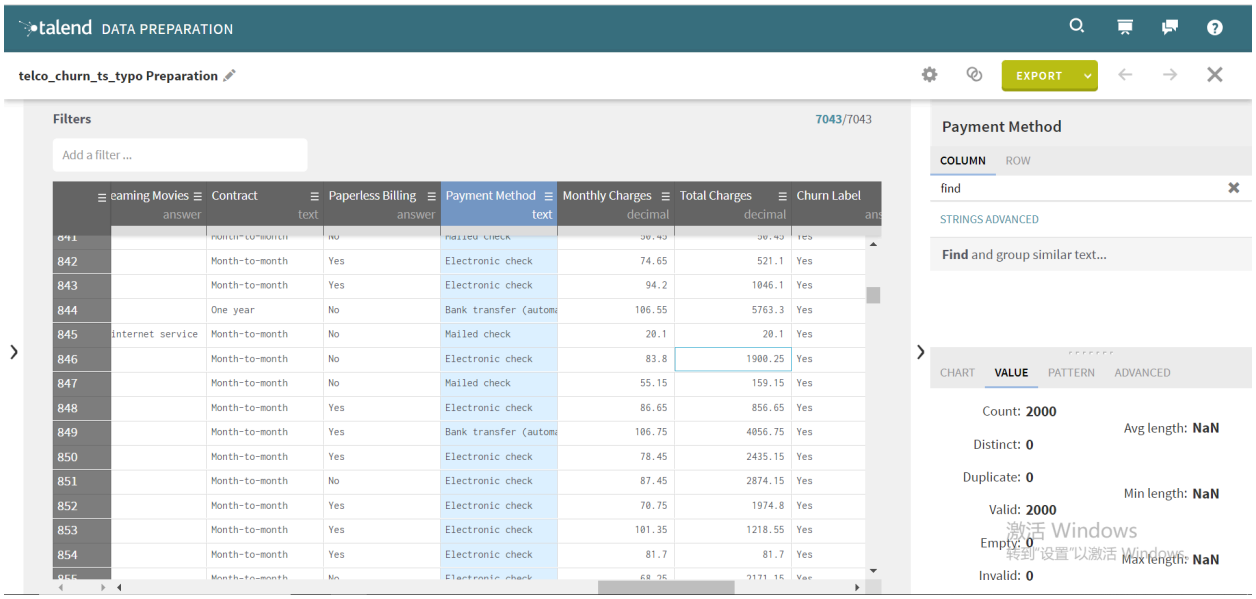
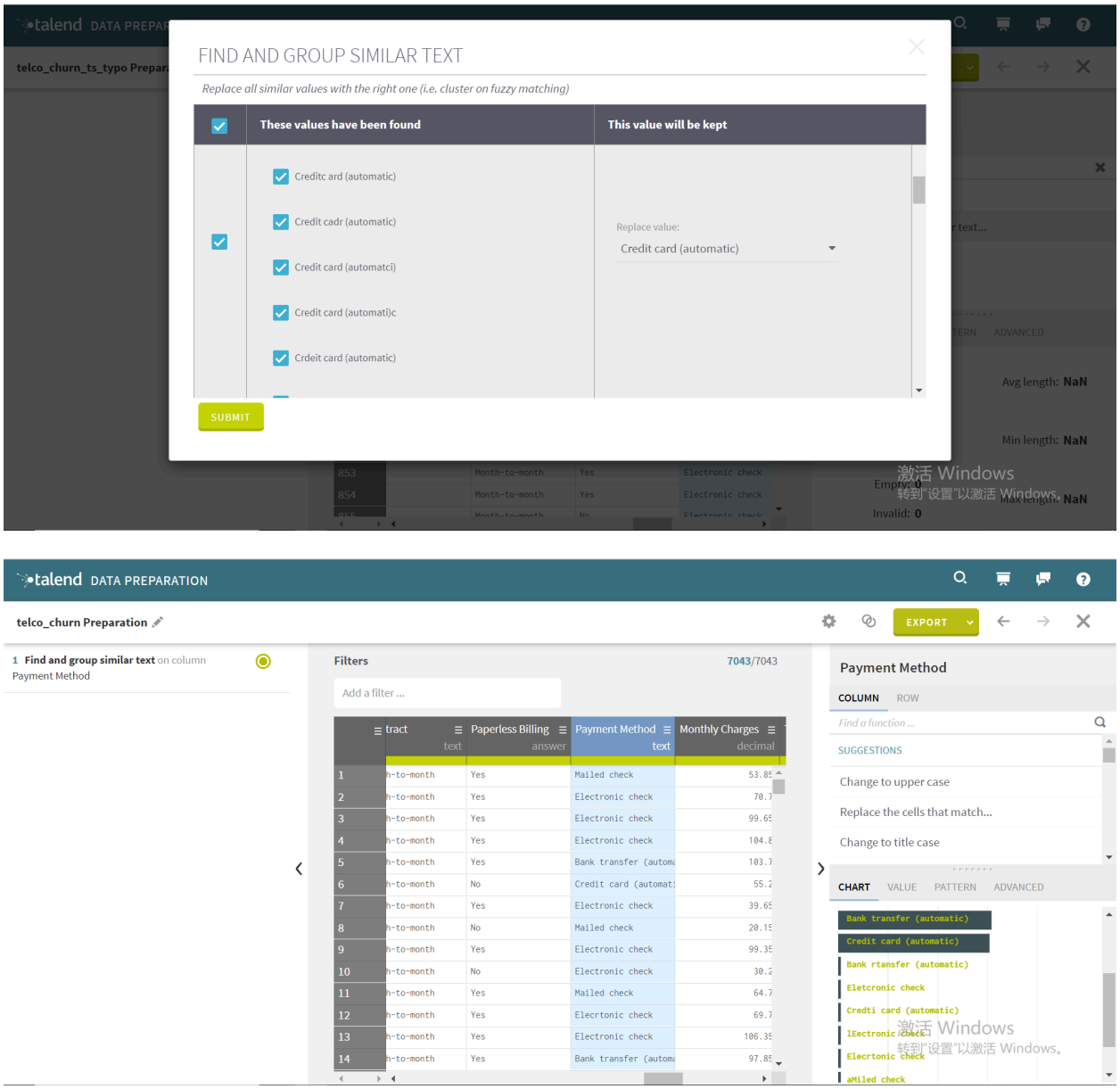
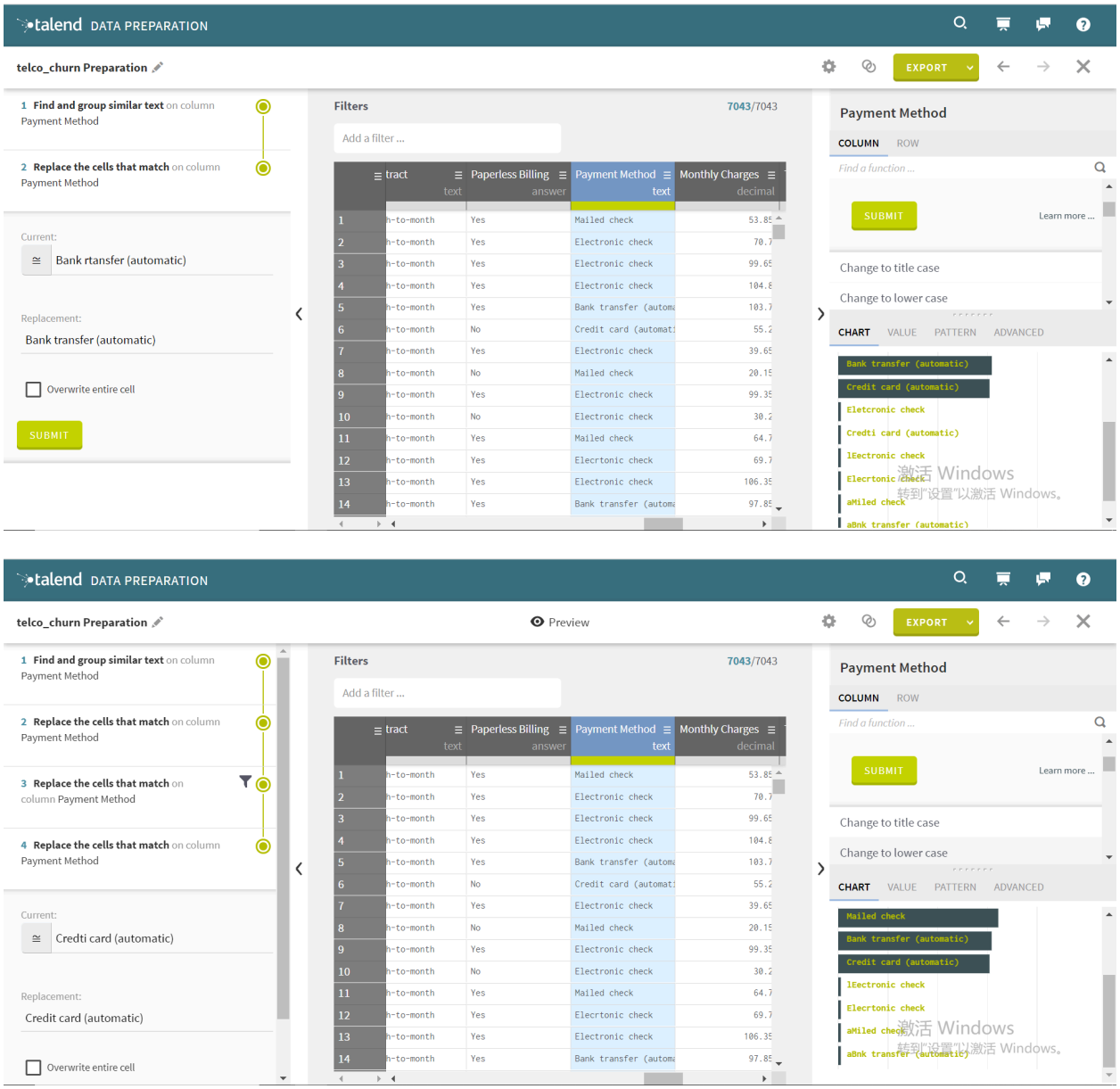


Figure 4.5.2&3: Using ‘FOUND AND GROUP SIMLIAR TEXT’ function



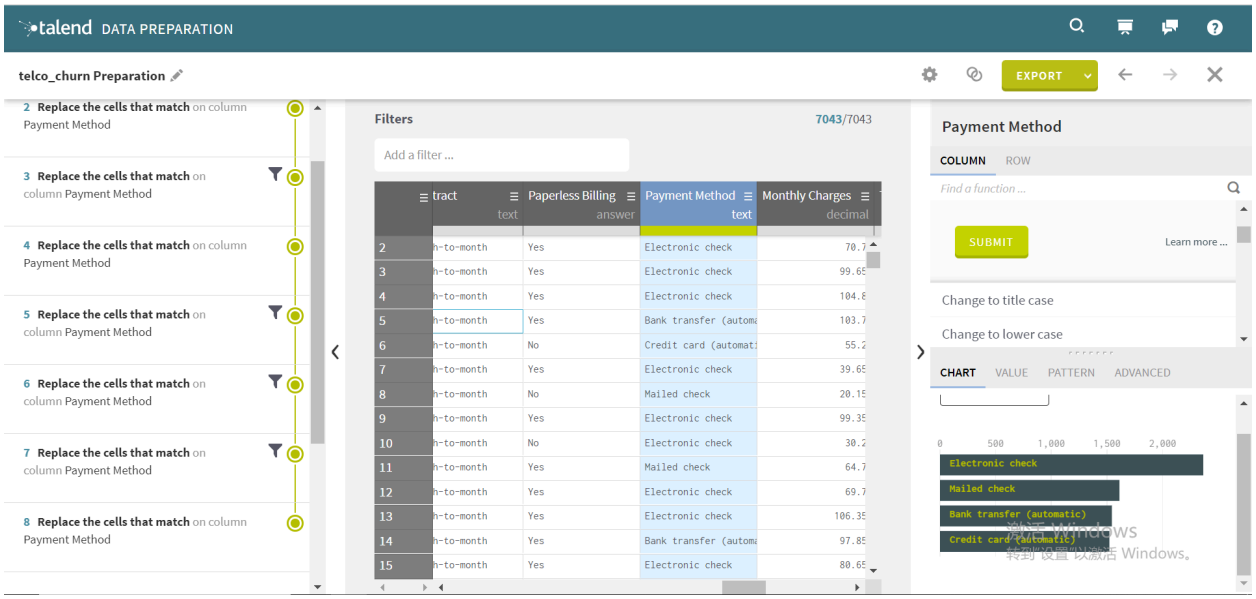
After using the 'FOUND AND GROUP SIMLIAR TEXT' function, there are still some misspelled values as shown in Figure 4.5.2&3. At this time, we need to use the 'REPLACE THE CELLS THAT MATH' function to manually modify them.

Figure 4.5.4&5: Using 'REPLACE THE CELLS THAT MATH' function



Using the 'REPLACE THE CELLS THAT MATCH' function, ‘Bank rtransfer (automatic)’ was replaced by Bank transfer (automatic)’, and ‘Credit card (automatic)’ was replaced by ‘Credit card (automatic)’, as shown in Figure 4.5.4&5.

Figure 4.5.6: Dataset after fixing spelling errors

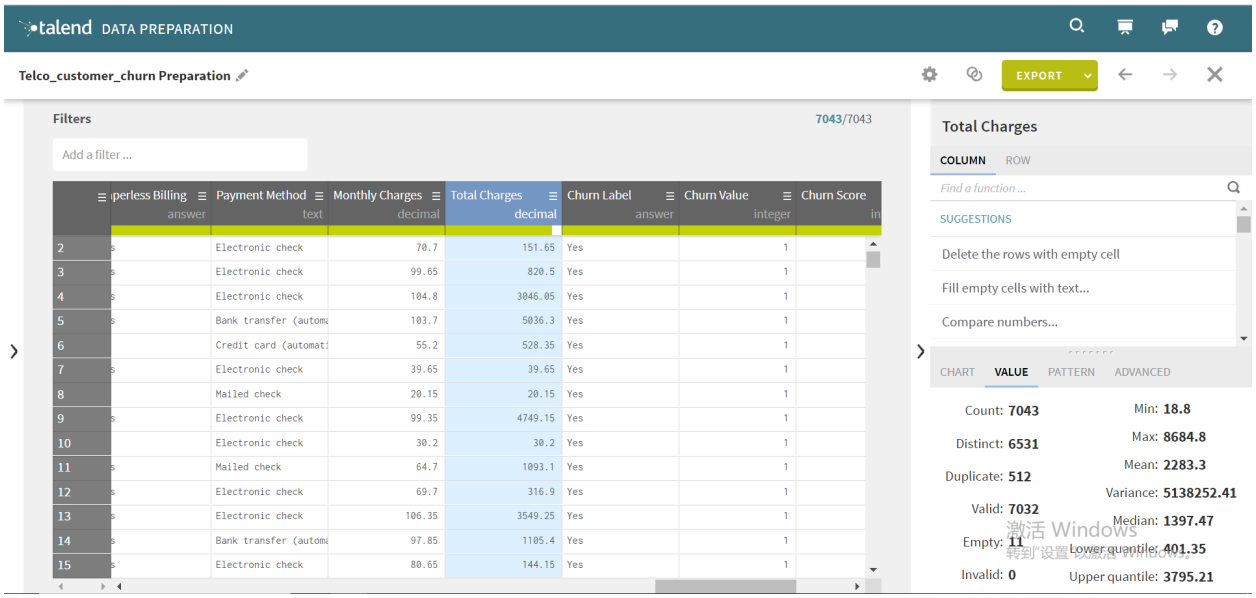


After modification, all spelling errors were corrected as shown in Figure 4.5.6.

4.5.2 Missing Value Imputation

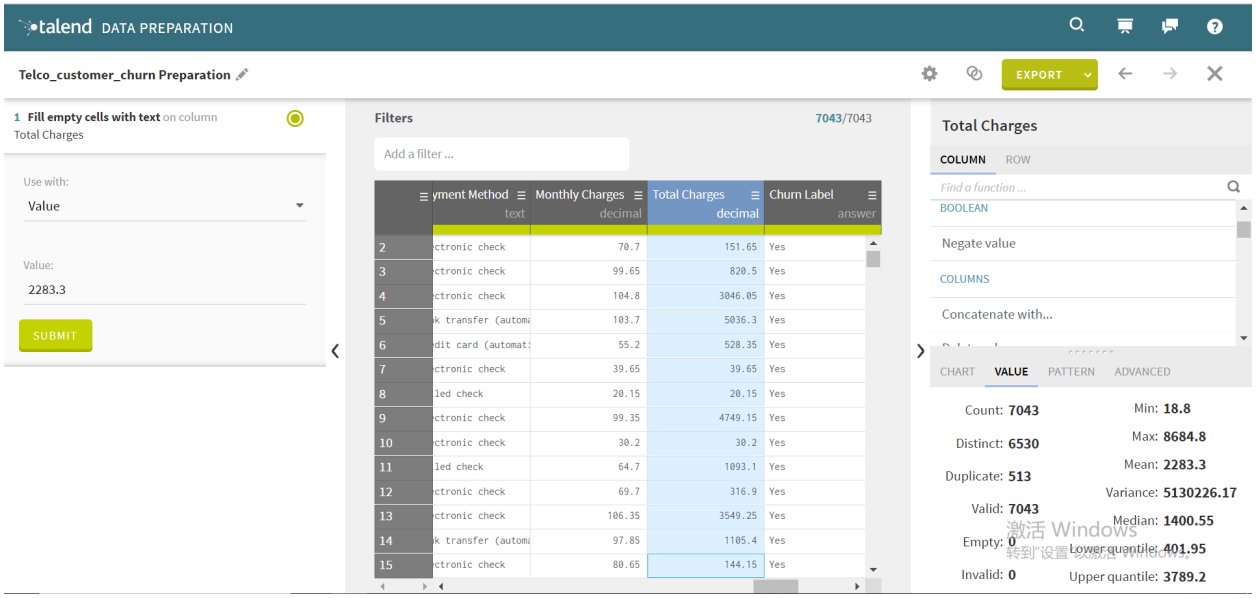
We found that there are 11 missing values in the 'Total Charges' column as shown in Figure 4.5.7.

Figure 4.5.7: Missing values in the 'Total Charges' column



We use the 'FILL EMPTY CELLS WITH TEXT' function to impute using the mean, the mean is 2283.3, so our missing value imputation is 2283.3 as shown in Figure 4.5.8.

Figure 4.5.8: Missing values imputation



After modification, all missing values are imputed as shown in Figure 4.5.8.

4.5.3 Feature Selection

Variables with low chi-squared values were removed to reduce noise in the model. Figure 4.5.9 shows the output of the Variable Selection Tool with the Chi-Square chosen as the selection criteria because this is a classification task. The input variables are CLTV, Contract, Dependents, Internet_Services, Monthly_Charges, Multiple_Lines, Onlinex_Security, Payment_Method, and Tenure_Months.

Figure 4.5.9: Variable selection

Variable Selection

Variable Name	Role	Measurement Level	Type	Label	Reasons for Rejection
CLTV	Input	Interval	Numeric		Varsel:Small Chi-square v...
Churn_Scor Variable Name	Rejected	Interval	Numeric	Churn Score	
Contract	Input	Nominal	Character		
Dependents	Input	Nominal	Character		
Device_Protection	Rejected	Nominal	Character	Device Protection	Varsel:Small Chi-square v...
Gender	Rejected	Nominal	Character		Varsel:Small Chi-square v...
Internet_Service	Input	Nominal	Character	Internet Service	
Latitude	Rejected	Interval	Numeric		Varsel:Small Chi-square v...
Longitude	Rejected	Interval	Numeric		Varsel:Small Chi-square v...
Monthly_Charges	Input	Interval	Numeric	Monthly Charges	Varsel:Small Chi-square v...
Multiple_Lines	Input	Nominal	Character	Multiple Lines	Varsel:Small Chi-square v...
Online_Backup	Rejected	Nominal	Character	Online Backup	Varsel:Small Chi-square v...
Online_Security	Input	Nominal	Character	Online Security	
Paperless_Billing	Rejected	Nominal	Character	Paperless Billing	Varsel:Small Chi-square v...
Partner	Rejected	Nominal	Character		Varsel:Small Chi-square v...
Payment_Method	Input	Nominal	Character	Payment Method	Varsel:Small Chi-square v...
Phone_Service	Rejected	Nominal	Character	Phone Service	Varsel:Small Chi-square v...
Senior_Citizen	Rejected	Nominal	Character	Senior Citizen	Varsel:Small Chi-square v...
Streaming_Movies	Rejected	Nominal	Character	Streaming Movies	Varsel:Small Chi-square v...
Streaming_TV	Rejected	Nominal	Character	Streaming TV	Varsel:Small Chi-square v...
Tech_Support	Rejected	Nominal	Character	Tech Support	Varsel:Small Chi-square v...
Tenure_Months	Input	Interval	Numeric	Tenure Months	
Total_Charges	Rejected	Interval	Numeric	Total Charges	Varsel:Small Chi-square v...

4.5.4 Normalization

Interval variables were normalized using log transformation, as shown in Figure 4.5.11.

Figure 4.5.10&11: Log transformation diagram and output

Computed Transformations
(maximum 500 observations printed)

Input Name	Role	Input Level	Name	Level	Formula
CLTV	INPUT	INTERVAL	LOG_CLTV	INTERVAL	log(CLTV + 1)
Monthly_Charges	INPUT	INTERVAL	LOG_Monthly_Charges	INTERVAL	log(Monthly_Charges + 1)
Tenure_Months	INPUT	INTERVAL	LOG_Tenure_Months	INTERVAL	log(Tenure_Months + 1)

4.6 Model (SEMMA: Step 4)

4.6.1 Importing Data

The cleaned dataset was imported into SAS Enterprise Miner using the File Import node, the steps are as follows.

Figure 4.6.1: SAS diagram

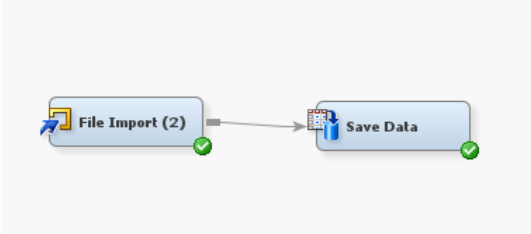
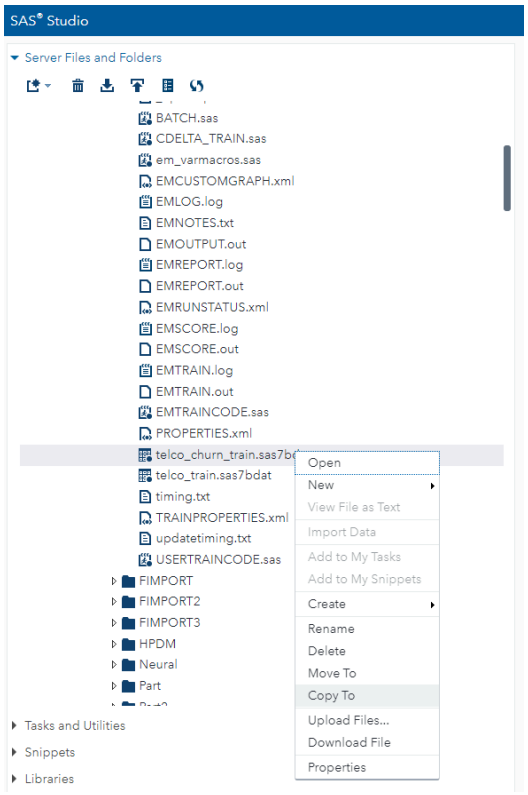


Figure 4.6.2: File import settings

Output Options	
Variables	
Filename Prefix	telco_churn
Replace Existing Files	Yes
All Observations	Yes
Number of Observations	1000

Then we get the destination path from Save Data results as follows:
/home/u63454162/7005_GP/Workspaces/EMWS2/EMSave/telco_churn_TRAIN.sas7bdat

Figure 4.6.3&4: Copying .sas7bdat file to sasuser.v94 folder.



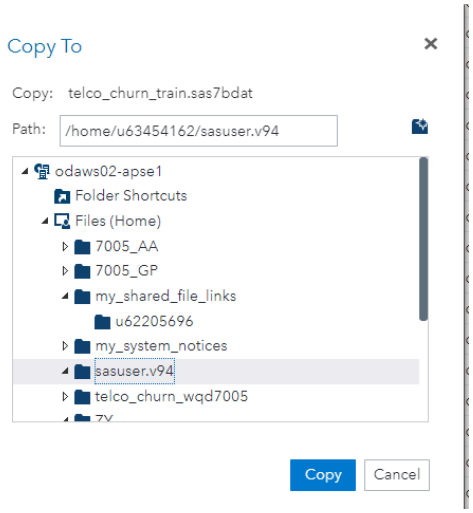
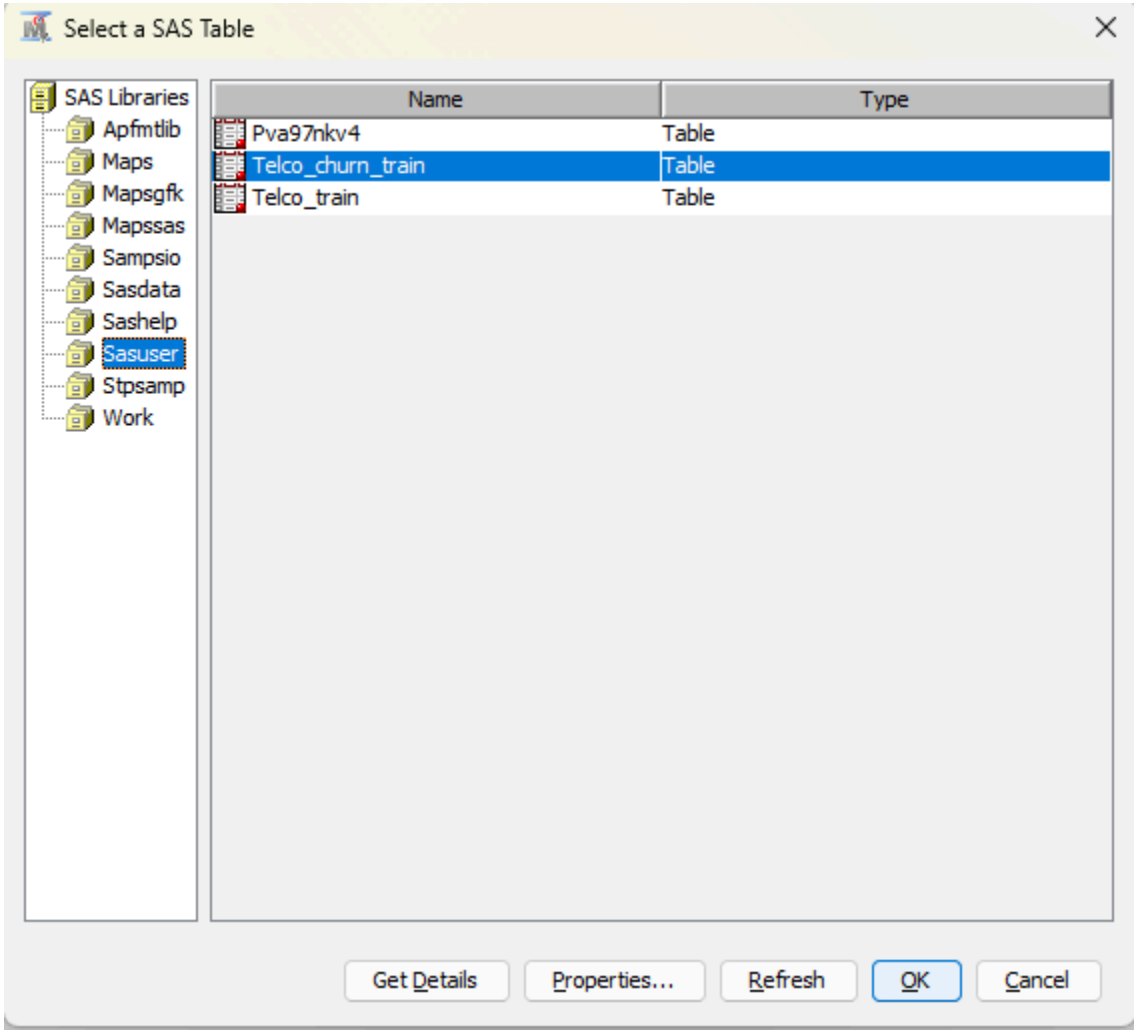


Figure 4.6.5&6: Creating SAS Data Source



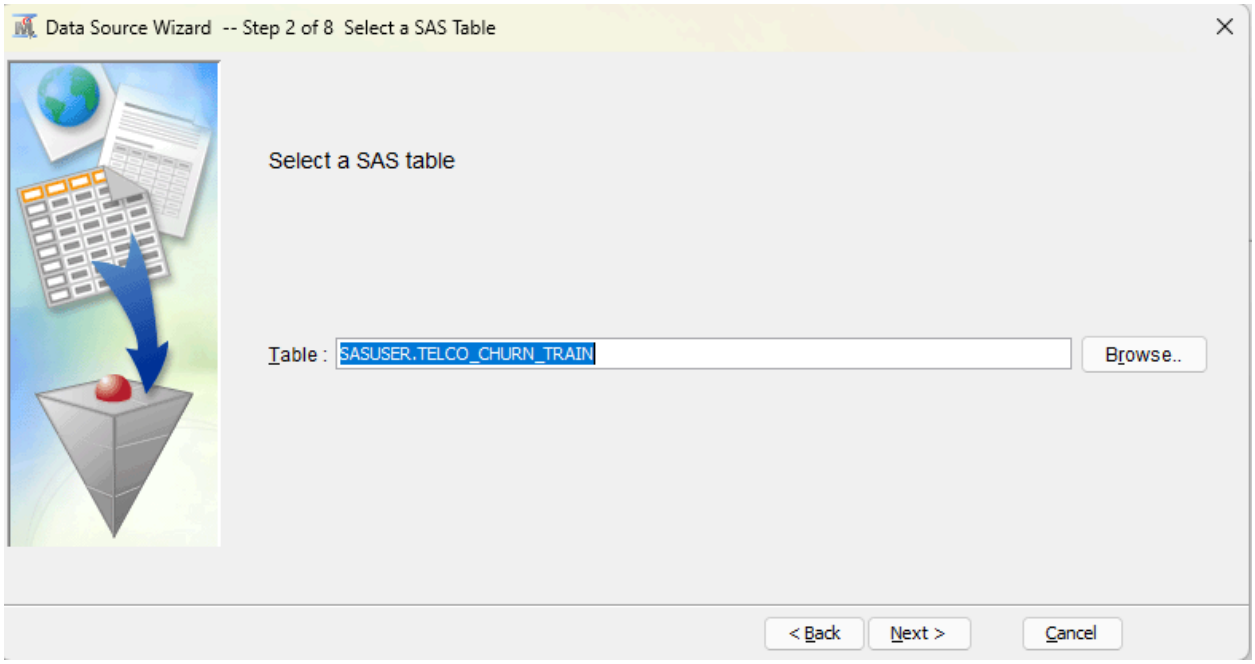


Figure 4.6.7: Data import settings.

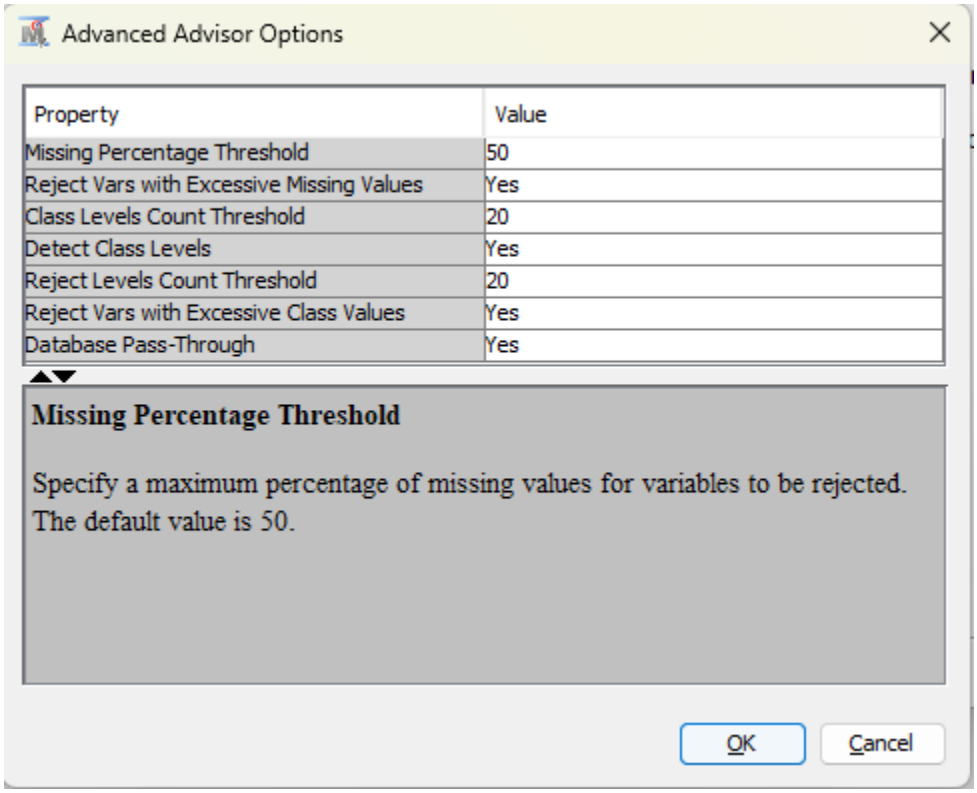
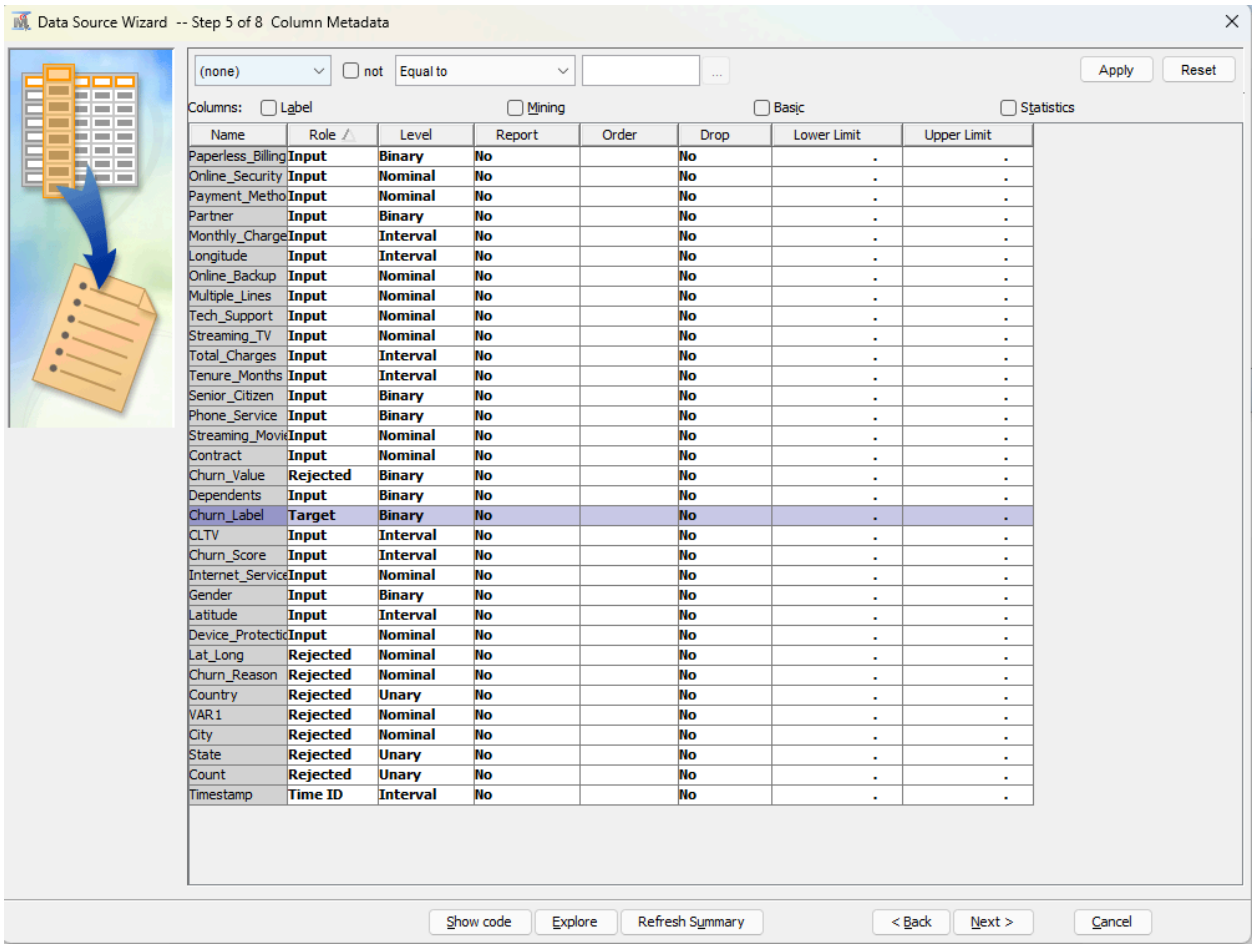


Figure 4.6.8: Defining Target variable role and checking variable levels.



Data Source Wizard -- Step 5 of 8 Column Metadata

(none) | not | Equal to | Apply | Reset

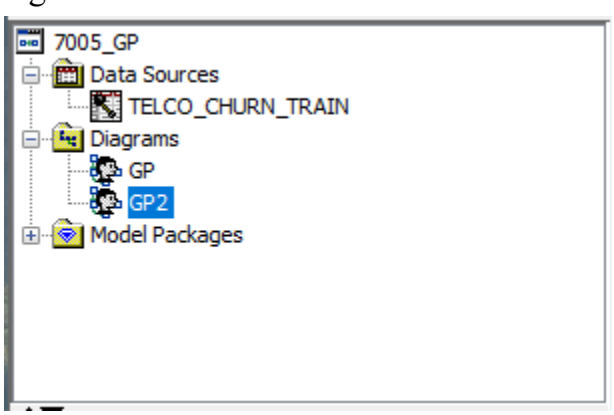
Columns: Label | Mining | Basic | Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Paperless_Billing	Input	Binary	No		No	.	.
Online_Security	Input	Nominal	No		No	.	.
Payment_Metho	Input	Nominal	No		No	.	.
Partner	Input	Binary	No		No	.	.
Monthly_Charge	Input	Interval	No		No	.	.
Longitude	Input	Interval	No		No	.	.
Online_Backup	Input	Nominal	No		No	.	.
Multiple_Lines	Input	Nominal	No		No	.	.
Tech_Support	Input	Nominal	No		No	.	.
Streaming_TV	Input	Nominal	No		No	.	.
Total_Charges	Input	Interval	No		No	.	.
Tenure_Months	Input	Interval	No		No	.	.
Senior_Citizen	Input	Binary	No		No	.	.
Phone_Service	Input	Binary	No		No	.	.
Streaming_Movi	Input	Nominal	No		No	.	.
Contract	Input	Nominal	No		No	.	.
Churn_Value	Rejected	Binary	No		No	.	.
Dependents	Input	Binary	No		No	.	.
Churn_Label	Target	Binary	No		No	.	.
CLTV	Input	Interval	No		No	.	.
Churn_Score	Input	Interval	No		No	.	.
Internet_Service	Input	Nominal	No		No	.	.
Gender	Input	Binary	No		No	.	.
Latitude	Input	Interval	No		No	.	.
Device_Protectio	Input	Nominal	No		No	.	.
Lat_Long	Rejected	Nominal	No		No	.	.
Churn_Reason	Rejected	Nominal	No		No	.	.
Country	Rejected	Unary	No		No	.	.
VAR1	Rejected	Nominal	No		No	.	.
City	Rejected	Nominal	No		No	.	.
State	Rejected	Unary	No		No	.	.
Count	Rejected	Unary	No		No	.	.
Timestamp	Time ID	Interval	No		No	.	.

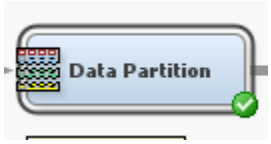
Show code | Explore | Refresh Summary | < Back | Next > | Cancel

Data Source is successfully created in SAS as shown in Figure 4.6.9.

Figure 4.6.9: Created data source



4.6.2 Data Partition



Prior to the modeling stage, we split the dataset to create both training and validation sets. The validation data is used to perform model selection from a group of candidate models. The training to validation split ratio for the dataset is 70:30 as shown in Figure 4.6.10.

Figure 4.6.10: Output of Data Partition

Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Churn_Label	.	No	5174	73.4630	Churn Label
Churn_Label	.	Yes	1869	26.5370	Churn Label

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Churn_Label	.	No	3621	73.4632	Churn Label
Churn_Label	.	Yes	1308	26.5368	Churn Label

Data=VALIDATE

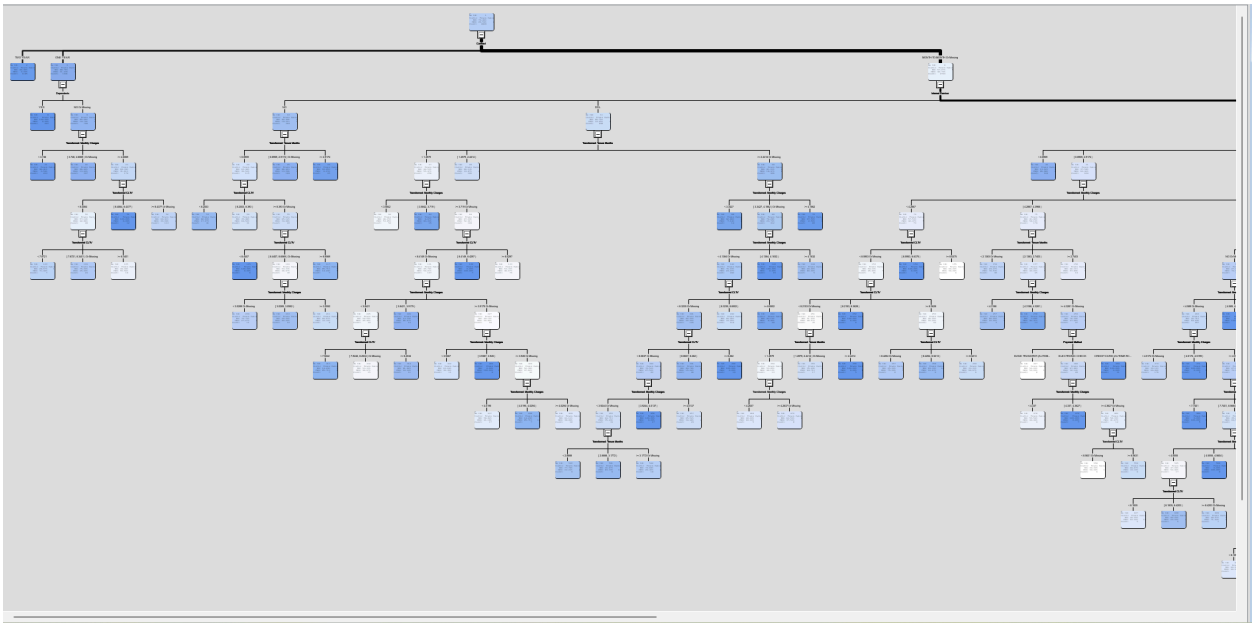
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Churn_Label	.	No	1553	73.4626	Churn Label
Churn_Label	.	Yes	561	26.5374	Churn Label

4.6.3 Decision Tree Model

Decision trees are a non-parametric supervised learning method used for classification and regression tasks. Its goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from data features. The topmost node is called the root node, and it represents the entire dataset. From this node, the data is split into subsets based on some characteristic condition, and these subsets form the "child node" of the root node. This process recurs, with each child node being further split based on other features until a stopping criterion is met.

In this project, hyperparameter tuning for decision tree models is conducted by varying Max Depth and Max Branches. Max Depth represents the maximum length from the root node to the last leaf node. This parameter is commonly used to control overfitting as a tree that is very deep may learn noise in the training data and result in poor performance on unseen data. Max Branches refers to the maximum number of splits allowed at each node. The performance of the model while varying these parameters will be discussed in the Assess section. Figure 4.6.11 shows one of the decision tree models.

Figure 4.6.11 Decision tree diagram



4.6.4 Gradient Boosting Model

Gradient boosting is a powerful ensemble machine learning technique that builds models in a sequential manner, where each new model progressively reduces the errors of the previous ones, typically using decision trees as weak learners. It involves optimizing a loss function, adding regularization, and using a learning rate to combine the predictions of individual trees into a single robust model. While it often achieves high predictive accuracy and can handle various types of data, it requires careful tuning of several hyperparameters to prevent overfitting and can be computationally intensive to train.

The same hyperparameters Max Depth and Max Branches were tested for the Gradient Boosting modes, and the results will be discussed in the Assess section.

4.6.5 Neural Network Model

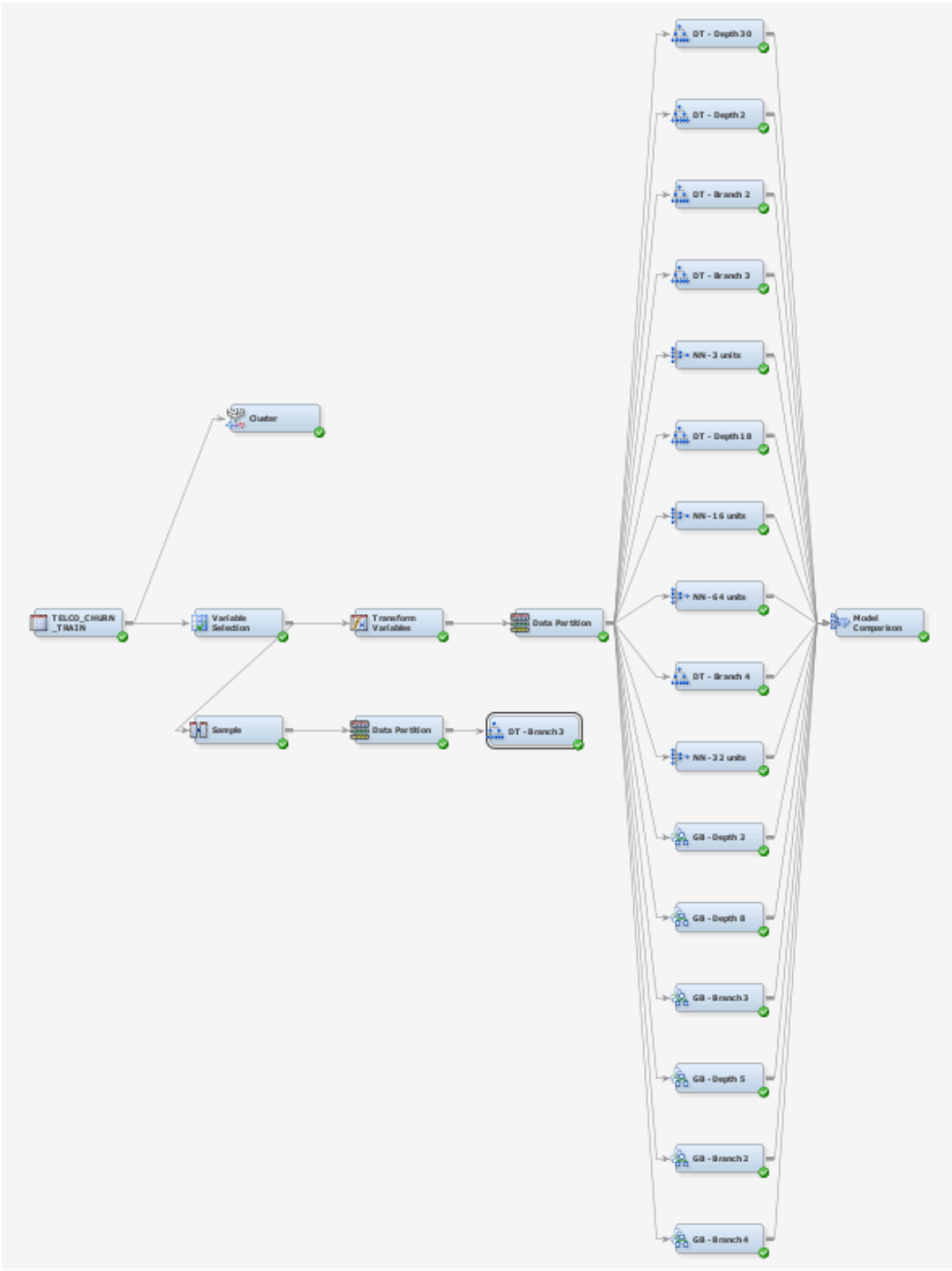
Neural networks are a class of machine learning algorithms modeled after the neural structures of the brain, designed to recognize patterns and interpret sensory data through a kind of machine perception, labeling, or clustering. They consist of layers of interconnected nodes or "neurons," with each layer transforming the input data into a more abstract and composite representation. Neural networks are particularly adept at handling complex, non-linear relationships within data and are widely used in applications ranging from image and speech recognition to natural language processing and autonomous vehicles, owing to their versatility and high performance in a wide range of complex tasks.

For Neural Network models, the number of hidden units was the hyperparameter that was tested. This parameter controls the complexity and number of parameters in a neural network. In a similar way to depth in decision trees, this parameter is commonly used to control overfitting in a neural network. The results of tuning this parameter will be discussed in the Assess section.

4.7 Assess (SEMMA: Step 5)

Figure 4.7.1 shows the whole SAS process and all the models we used. Model comparison is implemented to choose the best model, and oversampling is implemented to improve the result of imbalanced data.

Figure 4.7.1: Whole process diagram



4.7.1 Model Comparison

We chose misclassification rate of validation as the main criterion to select the best model. As shown in Figure 4.7.2, the selected model is Decision Tree with Branch 3, because the misclassification rate of validation is 0.174078, which is the lowest one among all models.

Figure 4.7.2: Model comparison using misclassification rate:

Selected Model	Predecessor Node	Model Node	Model Description ▲	Train: Misclassification Rate	Selection Criterion: Valid: Misclassification Rate
	Tree2	Tree2	DT - Branch 2	0.168594	0.178335
Y	Tree4	Tree4	DT - Branch 3	0.158247	0.174078
	Tree7	Tree7	DT - Branch 4	0.15561	0.18685
	Tree	Tree	DT - Depth 18	0.199229	0.201041
	Tree5	Tree5	DT - Depth 2	0.242443	0.226112
	Tree6	Tree6	DT - Depth 30	0.158247	0.174078
	Boost2	Boost2	GB - Branch 2	0.195577	0.181646
	Boost	Boost	GB - Branch 3	0.189085	0.182119
	Boost4	Boost4	GB - Branch 4	0.186245	0.18685
	Boost6	Boost6	GB - Depth 2	0.189085	0.182119
	Boost5	Boost5	GB - Depth 5	0.13025	0.202933
	Boost7	Boost7	GB - Depth 8	0.026983	0.199622
	Neural3	Neural3	NN - 16 units	0.1972	0.190634
	Neural2	Neural2	NN - 3 units	0.194157	0.18685
	Neural	Neural	NN - 32 units	0.199432	0.181646
	Neural4	Neural4	NN - 64 units	0.19436	0.186377

Figure 4.7.3 shows the classification result of each model, where we can find the classified number of False-Negative (predicted negative and actual positive), True-Negative (predicted negative and actual negative), False-Positive (predicted positive and actual negative), and True-Positive (predicted positive and actual positive).

Figure 4.7.3: Event classification table

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Boost7	GB - Depth 8	TRAIN	Churn_Label	Churn Label	105	3593	28	1203
Boost7	GB - Depth 8	VALIDATE	Churn_Label	Churn Label	269	1400	153	292
Boost6	GB - Depth 2	TRAIN	Churn_Label	Churn Label	639	3328	293	669
Boost6	GB - Depth 2	VALIDATE	Churn_Label	Churn Label	266	1434	119	295
Boost5	GB - Depth 5	TRAIN	Churn_Label	Churn Label	478	3457	164	830
Boost5	GB - Depth 5	VALIDATE	Churn_Label	Churn Label	297	1421	132	264
Boost	GB - Branch 3	TRAIN	Churn_Label	Churn Label	639	3328	293	669
Boost	GB - Branch 3	VALIDATE	Churn_Label	Churn Label	266	1434	119	295
Boost4	GB - Branch 4	TRAIN	Churn_Label	Churn Label	642	3345	276	666
Boost4	GB - Branch 4	VALIDATE	Churn_Label	Churn Label	271	1429	124	290
Boost2	GB - Branch 2	TRAIN	Churn_Label	Churn Label	650	3307	314	658
Boost2	GB - Branch 2	VALIDATE	Churn_Label	Churn Label	262	1431	122	299
Neural2	NN - 3 units	TRAIN	Churn_Label	Churn Label	655	3319	302	653
Neural2	NN - 3 units	VALIDATE	Churn_Label	Churn Label	275	1433	120	286
Tree	DT - Depth 18	TRAIN	Churn_Label	Churn Label	648	3287	334	660
Tree	DT - Depth 18	VALIDATE	Churn_Label	Churn Label	282	1410	143	279
Neural	NN - 32 units	TRAIN	Churn_Label	Churn Label	624	3262	359	684
Neural	NN - 32 units	VALIDATE	Churn_Label	Churn Label	249	1418	135	312
Tree5	DT - Depth 2	TRAIN	Churn_Label	Churn Label	506	2932	689	802
Tree5	DT - Depth 2	VALIDATE	Churn_Label	Churn Label	201	1276	277	360
Tree6	DT - Depth 30	TRAIN	Churn_Label	Churn Label	462	3303	318	846
Tree6	DT - Depth 30	VALIDATE	Churn_Label	Churn Label	220	1405	148	341
Neural3	NN - 16 units	TRAIN	Churn_Label	Churn Label	628	3277	344	680
Neural3	NN - 16 units	VALIDATE	Churn_Label	Churn Label	261	1411	142	300
Neural4	NN - 64 units	TRAIN	Churn_Label	Churn Label	604	3267	354	704
Neural4	NN - 64 units	VALIDATE	Churn_Label	Churn Label	248	1407	146	313
Tree2	DT - Branch 2	TRAIN	Churn_Label	Churn Label	534	3324	297	774
Tree2	DT - Branch 2	VALIDATE	Churn_Label	Churn Label	245	1421	132	316
Tree4	DT - Branch 3	TRAIN	Churn_Label	Churn Label	462	3303	318	846
Tree4	DT - Branch 3	VALIDATE	Churn_Label	Churn Label	220	1405	148	341
Tree7	DT - Branch 4	TRAIN	Churn_Label	Churn Label	455	3309	312	853
Tree7	DT - Branch 4	VALIDATE	Churn_Label	Churn Label	221	1379	174	340

Then we can calculate the result of Hyperparameter Tuning using metrics below, and the result is as shown in table 4.7.1.

For Decision Tree and Gradient Boosting models, increasing the Depth (Max Depth) seems to improve accuracy and F1 score in some cases, but this trend is not always consistent in the validation results, which may be a result of overfitting. The neural network model showed less difference between training and validation results, i.e.there is no significant overfitting, which may indicate that the model is more able to generalize to unseen data.

We also noticed that Decision Tree with depth set at 2 has the lowest accuracy and F1-score in training results, this could be the result of underfitting, indicating the classification model cannot learn data with such a shallow tree.

Since the dataset is imbalanced, we should focus on the F1-score instead of Accuracy to select the best model. The reason for this is that F1-score is calculated as the harmonic mean of precision and recall and hence gives a more balanced representation of the model performance at predicting both the minority and majority class instead of only the majority class. Thus the best model is Decision Tree with 3 branches and depth set at 30, which has the highest F1-score of 0.6471 in validation results.

- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $F1 = \frac{2 \times precision \times recall}{precision + recall}$
- $Accuracy = \frac{TP+TN}{TP+FN+TN+FP}$

Table 4.7.1: Results of Hyperparameter Tuning

Model	Hyperparameter	Value	Train Results		Validation Results			
			Accuracy	F1-Score	Accuracy	F1-Score	Precision	Recall
Decision Tree	Max Branches	2	0.8314	0.6507	0.8217	0.6264	0.7054	0.5633
		3	0.8466	0.6961	0.8240	0.6471	0.6917	0.6078
		4	0.8442	0.6855	0.8169	0.6311	0.6783	0.5900
	Max Depth	2	0.7576	0.5731	0.7739	0.6010	0.5651	0.6417
		18	0.8008	0.5734	0.7990	0.5677	0.6611	0.4973
		30	0.8466	0.6961	0.8240	0.6471	0.6917	0.6078
Gradient Boosting	Max Branches	2	0.8044	0.5772	0.8184	0.6090	0.7102	0.5330
		3	0.8144	0.5974	0.8155	0.6022	0.7014	0.5276
		4	0.8152	0.5974	0.8127	0.5909	0.7027	0.5098
	Max Depth	2	0.8144	0.5974	0.8155	0.6022	0.7014	0.5276
		18	0.8805	0.7503	0.8013	0.5682	0.6691	0.4938
		30	0.9444	0.8881	0.7966	0.5639	0.6541	0.4955
Neural Network	Number of Hidden Units	3	0.8038	0.5855	0.8160	0.6137	0.6928	0.5508
		16	0.8038	0.5915	0.8198	0.6231	0.7000	0.5615
		32	0.8065	0.5968	0.8127	0.6095	0.6821	0.5508
		64	0.8058	0.6031	0.8198	0.6297	0.6923	0.5775

4.7.2 Variable Importance Analysis

Figure 4.7.4 shows the results of variable importance analysis for Decision Tree model (Branch 3). Variable importance is a method in machine learning models to evaluate the relative contribution of features in the decision of the prediction model. Several measures of the different variables and their importance are listed here:

Number of Splitting Rules: Number of splitting rules using this feature in the model.
Importance: The importance of a feature, usually based on some criterion such as how much it reduces model error.
Validation Importance: The feature importance computed on the validation dataset.
Ratio of Validation to Training Importance: Ratio of validation to training importance.

The following conclusions can be drawn from this figure:

Contract is the most important variable, having the highest importance of 1.0000 on both the training and validation sets. This means that the contract type has the greatest impact on the prediction of the model.

The LOG_Tenure_Months and LOG_Monthly_Charges variables were also heavily used for splitting rules (13 and 25 times, respectively), showing that they have a relatively large impact on the model's predictions, although they are less important than the contract type.

LOG_CLTV, although used frequently in model splitting (22 times), has a low importance score (0.4758) and a reduced importance score of 0.2345 on the validation set, suggesting that it may have a tendency to overfit on the training data.

Internet_Service has a large difference in importance between training and validation, and its importance on the validation set is 1.1765 times that of the training set, indicating that this variable contributes more to the model prediction on the validation set.

Dependents and Payment_Method have low importance scores, but their validation set importance is close to that of the training set (ratio is close to 1), which may indicate that these variables contribute steadily to the generalization ability of the model.

Online_Security and Multiple_Lines have little to no contribution (importance 0) in this particular branch of the decision tree, which could mean that they have very little impact on the prediction, or that they are not critical factors in the decision in this branch of the model.

Figure 4.7.4: DT - Branch 3 Variable Importance

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Contract		1	1.0000	1.0000	1.0000
LOG_Tenure_Months	Transforme...	13	0.7341	0.5936	0.8087
LOG_Monthly_Charges	Transforme...	25	0.5844	0.3184	0.5447
Internet_Service	Internet Ser...	1	0.5320	0.6260	1.1765
LOG_CLTV	Transforme...	22	0.4758	0.2345	0.4930
Dependents		3	0.3257	0.4151	1.2746
Payment_Method	Payment M...	2	0.1178	0.1182	1.0035
Online_Security	Online Sec...	1	0.1018	0.0924	0.9084
Multiple_Lines	Multiple Lin...	0	0.0000	0.0000	.

4.7.2 Oversampling

To improve the model performance, oversampling is implemented on the imbalanced data to balance the dataset by increasing the number of samples in the minority class, so that the model can better learn and recognize the minority class. The dataset is balanced to 50:50 as shown in Figure 4.7.6.

Figure 4.7.5: SAS diagram for oversampling

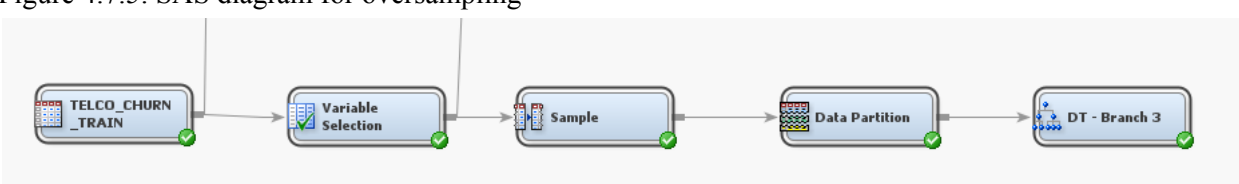


Figure 4.7.6: Oversampling details

Data=SAMPLE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Churn_Label	.	No	1869	50	Churn Label
Churn_Label	.	Yes	1869	50	Churn Label

As shown in Figure 4.7.7, the misclassification rate of validation is 0.184432 which is higher than that without validation (0.174078), thus oversampling does not improve the model performance. The possible reason could be overfitting. Oversampling can lead to a

model that fits too perfectly on the training data, especially if the oversampling method simply replicates minority class samples. This makes the model less able to generalize to real-world data.

Figure 4.7.7: Result of Oversampling

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Churn_Label	Churn Label	_NOBS_	Sum of Frequencies	2614.531	1123.469
Churn_Label	Churn Label	_MISC_	Misclassification Rate	0.133436	0.184432
Churn_Label	Churn Label	_MAX_	Maximum Absolute Err...	0.972964	1
Churn_Label	Churn Label	_SSE_	Sum of Squared Errors	513.89	351.436
Churn_Label	Churn Label	_ASE_	Average Squared Error	0.098276	0.156407
Churn_Label	Churn Label	_RASE_	Root Average Squared...	0.31349	0.395483
Churn_Label	Churn Label	_DIV_	Divisor for ASE	5229.061	2246.939
Churn_Label	Churn Label	_DFT_	Total Degrees of Free...	2614.531	.

Chapter 5: Conclusion and Recommendations

5.1 Summary of Insights

To conclude this report, the following summaries were drawn from the analyses conducted with respect to each objective formed earlier.

Table 5.1: Summary of Insights from the Explore Phase

Objective	**Summary**
To explore the relationship between the variables and the target variable, customer churn.	Contract is the most important variable to predict customer churn. The LOG_Tenure_Months and LOG_Monthly_Charges variables are secondly important. A longer contract, less monthly charges and longer tenure months would lead to less churn.
To build a model to predict customer churn using selected features.	Decision Tree, Gradient Boosting, and Neural Network models were built to predict customer churn, and different

	hyperparameters are implemented on the models.
To assess the model using reasonable metrics.	The models are assessed compared using metrics of accuracy, F1-score, precision, and recall. Since the dataset is imbalanced, the best-performed model is selected based on F1-score, it is Decision Tree with 3 branches and depth set at 30, which has the highest F1-score of 0.6471 in validation results.

5.2 Recommendations

- Incentivize Longer Contracts: Develop customer retention strategies that encourage longer contract commitments, such as discounted rates for extended contract periods, loyalty benefits, or graduated benefits that increase over the contract duration.

- Review Pricing Strategies: Since lower monthly charges are associated with reduced churn, consider reviewing and possibly restructuring pricing plans to offer more competitive rates, or introduce a tiered pricing system that rewards customer loyalty with lower monthly costs over time.

- Loyalty Programs: Implement or enhance loyalty programs that recognize and reward long-term customers. This could include exclusive offers, additional services, or discounts that become more substantial as the tenure with the company increases.

5.3 Reflections

In this case study, we conducted predictive modeling for customer churn. The imbalanced dataset posed a significant challenge as it is known to cause bias in models towards the majority class, leading to poor performance in churn prediction. To address this, strategies such as oversampling, undersampling, or specialized algorithms are essential. Additionally, the process of feature selection is also important. However, it requires a deep understanding of the business context to identify which attributes are most predictive of churn. This often involves applying statistical methods and predictive modeling to determine the significance of different features. Moreover, hyperparameter tuning stands out as a significant hurdle due to its complexity and the exhaustive nature of the search for the optimal parameter. Hyperparameter tuning also requires a lot of time and computational resources.

In this case study, we learned to apply the SEMMA methodology as well as new concepts such as clustering, time series analysis and association analysis. Although these techniques could not be applied to the customer churn dataset due to limitations of the data structure of our selected, through experimentation and the process of trying to apply these techniques, we have learned the key concepts behind the techniques.

For future research, there are several promising areas for further exploration. Advanced sampling techniques like Synthetic Minority Over-sampling Technique (SMOTE) or Adaptive Synthetic Sampling (ADASYN) offer more sophisticated ways to address data imbalance and could enhance model performance more than the random oversampling applied in this case study. Additionally, more recent ensemble methods that are specialized in imbalanced data could also be applied to improve churn prediction.

Appendix

- **Presentation Link:**

  Presentation video of our group project uploaded to YouTube.

  Below is our link:

  https://www.youtube.com/watch?v=8g_lwciHwtE

- **Teamwork & Collaboration:**

Name	**Contribution**	**Peer Evaluation**
TNEU ZHEN YANG	Sample, Explore, Model, Assess	Proficient in SAS operation and SEMMA process, and willing to help team members solve difficulties in the operation process. Excellent leadership skills, led the team to successfully complete the project.
HAN DIE	Introduction, Business Understanding, Dataset Description, Modify, Report Organization	Clear thinking and good at business understanding, context description, and data cleaning. Helped team members improve time management so that the team completed the work ahead of time.
SIYU JIANG	Data Integration, Conclusion and Recommendations, Result Description & Report Organization	Taking a serious attitude to pursue perfection and strive to learn new skills like Talend Data Integration. Helped the team to check and fill the gaps and improve the quality of the project report