Analyzing E-commerce Customer Behavior for Churn Prediction

Chapter 1: Introduction

This report embarks on a comprehensive analysis of e-commerce customer behavior, emphasizing the process and discussion of results obtained through sophisticated data mining techniques. In today's dynamic digital marketplace, understanding customer behavior is pivotal to address the challenge of churn and enhance customer retention strategies. The primary aim of this analysis is to leverage advanced data mining methodologies to explore an e-commerce dataset, predict customer churn, and derive actionable insights. The report concentrates on delineating the process, findings, and discussions stemming from the analysis.

Chapter 2: Dataset

2.1 Data Source and Origin

The dataset utilized in this analysis originates from Kaggle (https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction). It encompasses customer transaction records from an e-commerce platform and is relevant to understanding customer behavior within the online retail space.

2.2 Dataset Dimension

- Observations: The dataset contains a total of 5630 observations, each representing a unique customer transaction or interaction.
- Attributes: The dataset comprises 20 distinct attributes that encapsulate various facets of customer behavior and transactional information.

2.3 Dataset Attributes

The dataset encompasses a mix of categorical, numerical, and textual attributes, enabling a holistic understanding of customer behavior.

2.3.1 Unique Identifier: CustomerID

An integer-based identifier, ensuring each record corresponds to a unique customer, pivotal for tracking individual customer activities.

2.3.2 Target Variable: Churn

A categorical variable indicating whether a customer stopped purchasing (1 for churned, 0 for active), crucial for predictive analysis and understanding customer retention.

2.3.3 Additional Descriptors

- **Tenure**: Numerical attribute representing the duration of a customer's association with the platform.
- **PreferredLoginDevice**: Categorical attribute denoting the customer's preferred device for logging in.
- **CityTier**: Categorical attribute classifying customers based on city tiers.
- **WarehouseToHome**: Numerical attribute indicating the distance between the customer's home and the nearest warehouse.
- **PreferredPaymentMode**: Categorical attribute specifying the preferred payment mode of customers.
- **Gender**: Categorical attribute denoting the gender of customers.
- **HourSpendOnApp**: Numerical attribute representing the time spent by customers on the platform's application.
- **NumberOfDeviceRegistered**: Numerical attribute indicating the count of devices registered per customer.
- **PreferredOrderCat**: Categorical attribute reflecting the preferred order category of customers.
- **SatisfactionScore**: Numerical attribute indicating the satisfaction score of customers regarding the platform's services.
- **MaritalStatus**: Categorical attribute denoting the marital status of customers.
- **NumberOfAddress**: Numerical attribute specifying the count of addresses registered per customer.
- **Complain**: Binary attribute indicating whether a customer has raised any complaints in the last month.
- **OrderAmountHikeFromlastYear**: Numerical attribute indicating the percentage increase in order amount from the previous year.
- **CouponUsed**: Numerical attribute indicating the count of coupons used by customers in the last month.
- **OrderCount**: Numerical attribute denoting the count of orders placed by customers in the last month.
- **DaySinceLastOrder**: Numerical attribute indicating the number of days since the customer's last order.

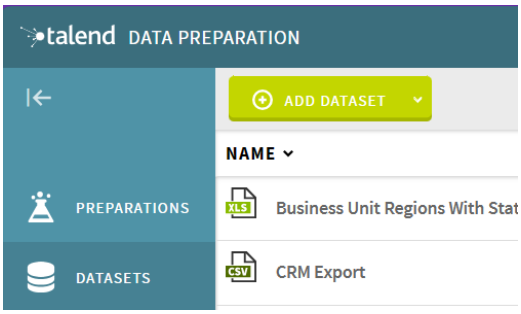
- **CashbackAmount**: Numerical attribute representing the average cashback amount received by customers in the last month.

Chapter 3: Methodology

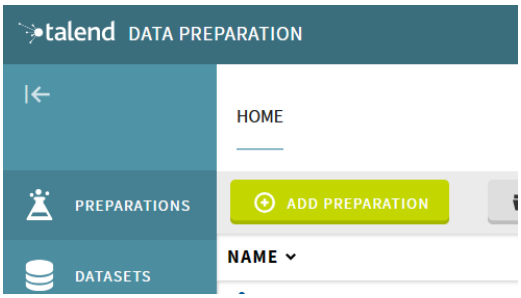
3.1 **Data Exploration**

The primary phase involved an exhaustive examination of each attribute within the dataset using Talend Data Preparation. This meticulous review facilitated a clearer understanding of the dataset's attributes and their potential impact on forecasting churn behavior within the e-commerce platform.

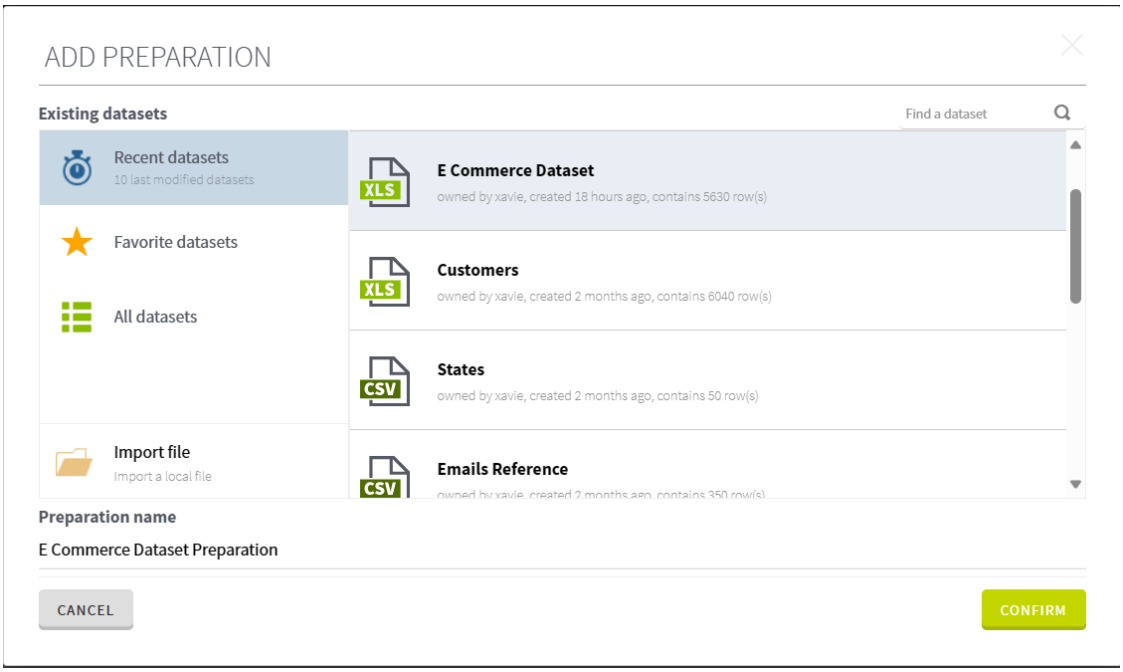
- First, add our raw dataset into Talend Data Preparation, under the "Datasets" tab



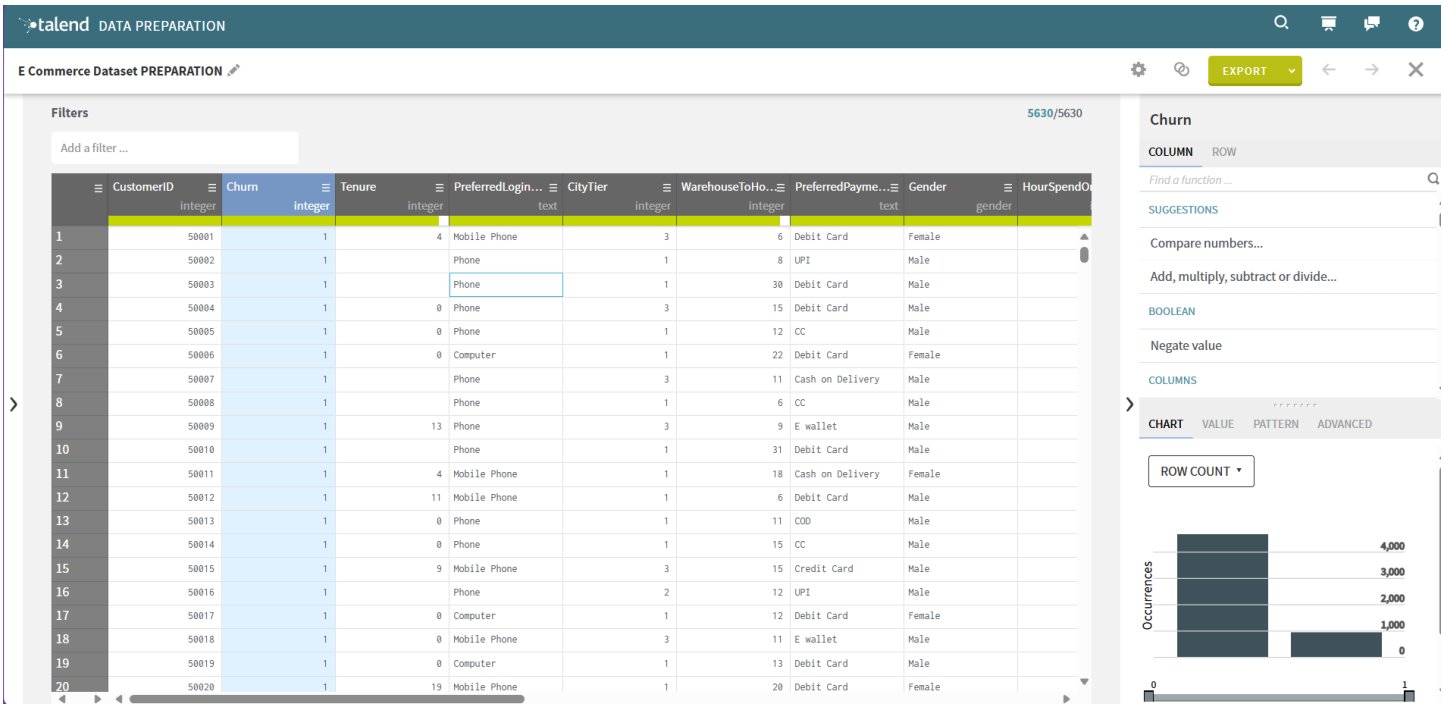
- Switch to the "Preparations" tab



- Then, select the dataset that we have added into Talend Data Preparation earlier



- Lastly, we could explore each column of the raw dataset by clicking on it

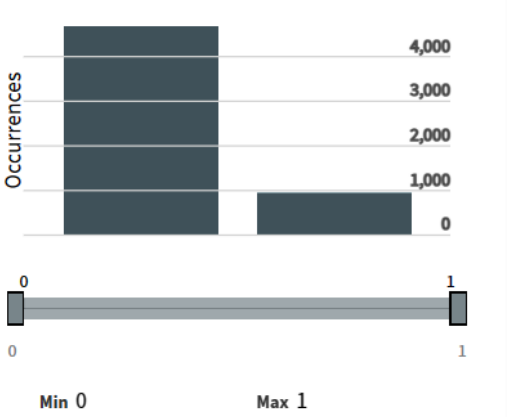


3.1.1 CustomerID: Unique Identifier

The CustomerID attribute was identified as a pivotal component, serving as a unique identifier for individual customers. Its significance lies in enabling precise individual-level analysis, allowing tracking and assessment of distinct customer behaviors and interactions within the platform.

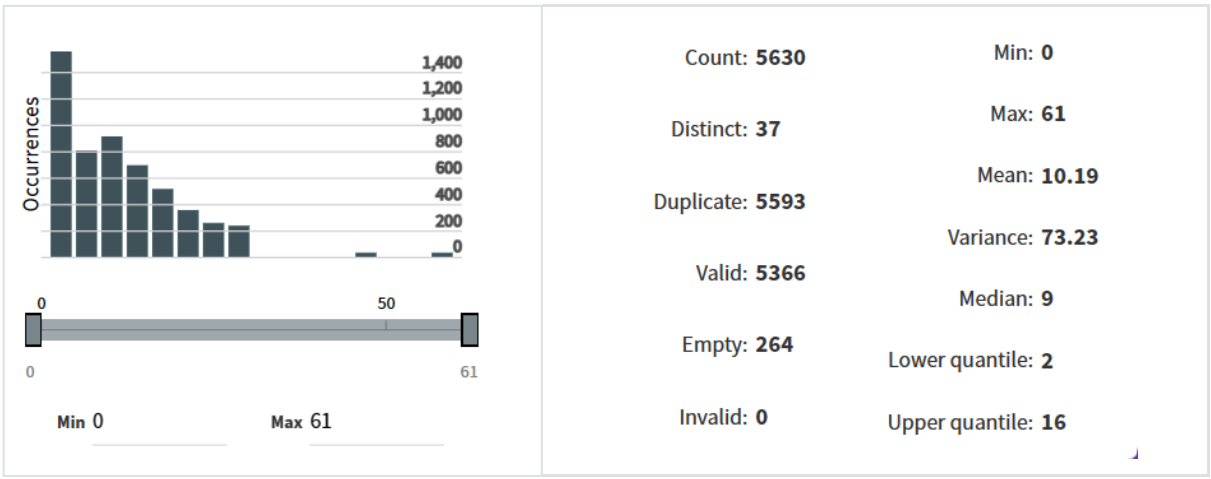
3.1.2 Churn: Target Variable

Churn emerged as the crucial target variable, signifying whether customers churned (1) or were retained (0) within the platform. This attribute forms the cornerstone for predictive modeling, ascertaining the propensity of customers to discontinue their engagement with the platform.



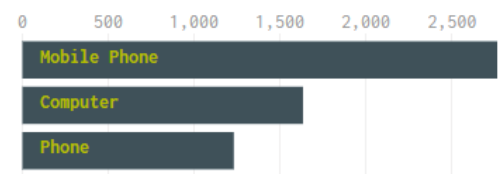
3.1.3 Tenure

Tenure provided valuable insights into the duration of customer association with the e-commerce platform. However, its integrity was compromised by the presence of 264 missing values, which we will impute with the mean value of 10. Despite this shortfall, the attribute holds significant potential to discern the relationship between customer longevity and churn propensity.



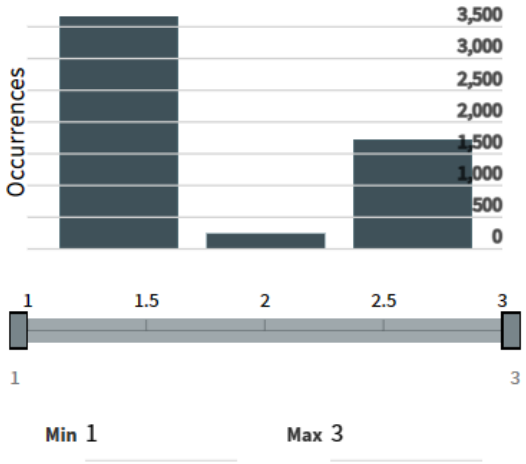
3.1.4 PreferredLoginDevice

The PreferredLoginDevice attribute furnished insights into customers' preferred devices for accessing the platform, which are Mobile Phone, Computer and Phone. It was identified that merging similar categories within this attribute could substantially enhance classification accuracy. This step aimed to streamline and consolidate device preferences for more effective modeling.



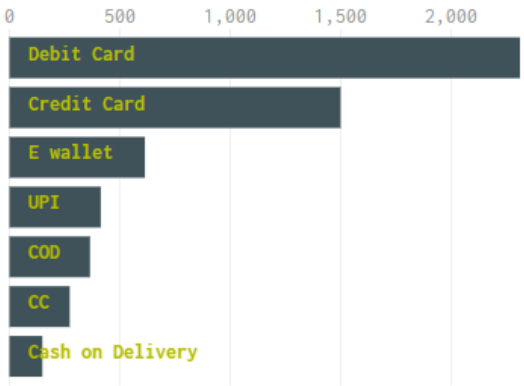
3.1.5 CityTier

The CityTier attribute offered insights into the classification of customers based on city tiers. Analyzing this attribute shed light on the distribution of customers across different city tiers, providing contextual information for understanding regional disparities in customer behavior and potential implications on churn.



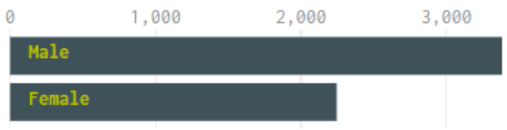
3.1.6 PreferredPaymentMode

The PreferredPaymentMode attribute encapsulates diverse payment preferences exhibited by customers during transactions on the e-commerce platform. This attribute offers critical insights into customers' preferred methods of payment, which are instrumental in tailoring payment strategies to enhance customer experience and retention. Similar to the insights derived from PreferredLoginDevice, it was discovered that certain categories within PreferredPaymentMode could be streamlined to enhance classification accuracy during modeling.



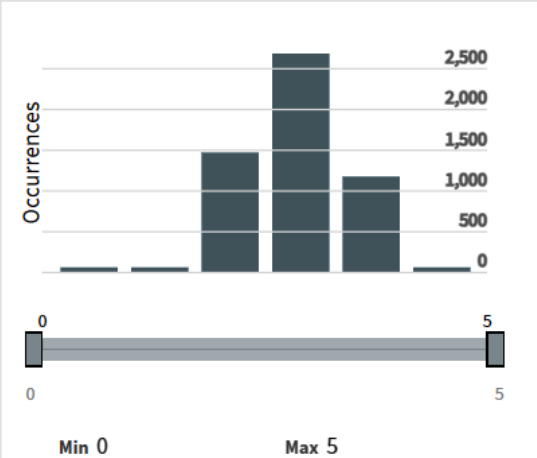
3.1.7 Gender

The Gender attribute provides insights into the gender distribution among customers. Analyzing gender-specific behaviors might uncover gender-centric patterns affecting churn, contributing to more targeted marketing and service strategies.



3.1.8 HourSpendOnApp

HourSpendOnApp reveals the time duration customers spent on the platform's application or website. Analyzing this attribute's missing values discovered during data exploration (255 observations) becomes crucial to understand correlations between customer engagement levels and their propensity to churn. These missing entries, similar to Tenure, were handled through mean imputation with the calculated value of 3. This approach ensures the dataset's completeness without compromising insights into customer engagement patterns.

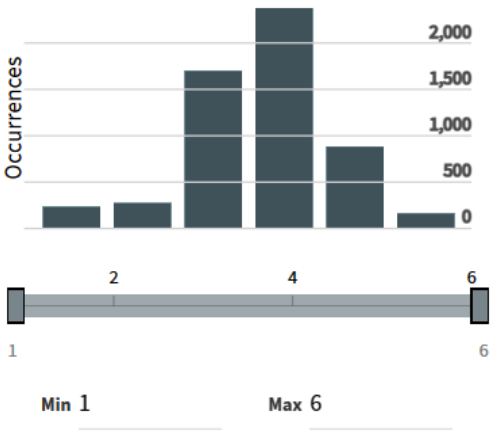


Count: **5630**
Distinct: **7**
Duplicate: **5623**
Valid: **5375**
Empty: **255**
Invalid: **0**

Min: **0**
Max: **5**
Mean: **2.93**
Variance: **0.52**
Median: **3**
Lower quantile: **2**
Upper quantile: **3**

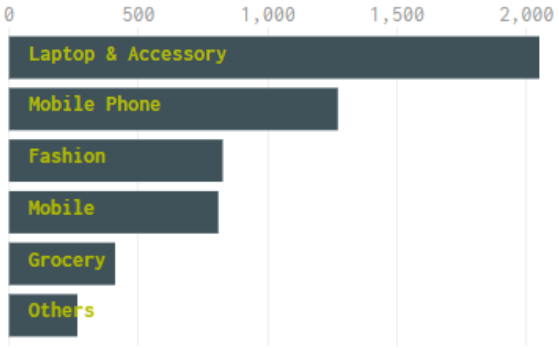
3.1.9 PreferredOrderCat

PreferedOrderCat captures customers' preferences for specific order categories. Analyzing these preferences might unveil category-specific behaviors impacting churn, enabling tailored strategies to enhance retention in popular order categories.



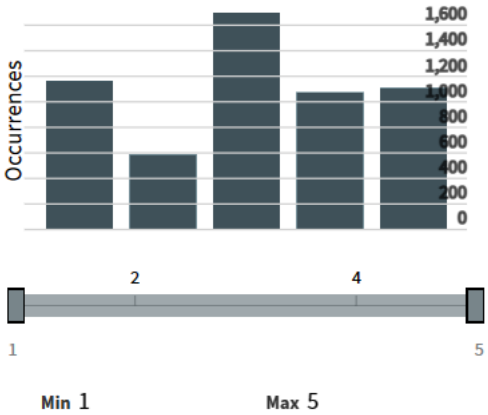
3.1.10 SatisfactionScore

The SatisfactionScore attribute denotes customers' satisfaction levels with the platform's services. Analysis of this attribute might highlight the relationship between service quality and customer retention.



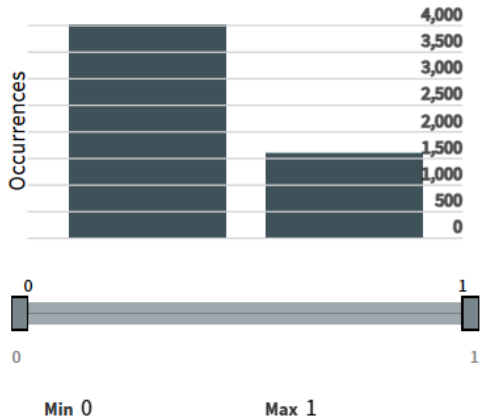
3.1.11 MaritalStatus

MaritalStatus reflects the marital status distribution among customers. Understanding if marital status influences churn behavior can assist in designing retention strategies tailored to specific demographic segments.



3.1.12 Complain

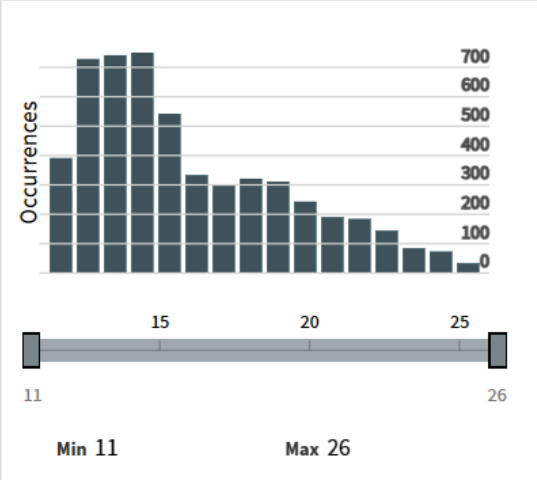
Complain indicates whether customers raised any complaints in the last month: 0 represents no, 1 represents yes. Analyzing this attribute might reveal the impact of customer grievances on churn rates and the importance of effective complaint resolution in retention efforts.



3.1.13 OrderAmountHikeFromlastYear

OrderAmountHikeFromlastYear denotes the percentage increase in order amounts from the previous year. This attribute encountered 265 missing values during exploration. Mean

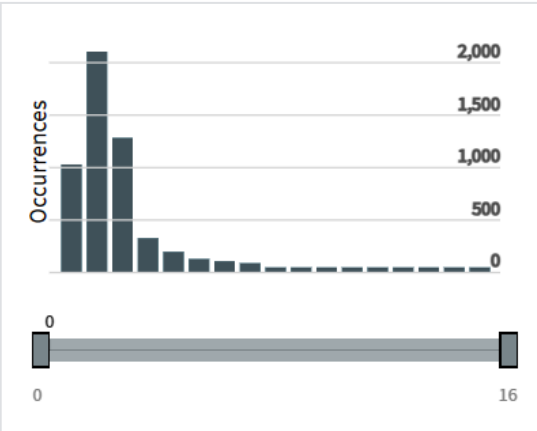
imputation using the calculated mean value of 16 was employed to ensure the attribute's representation of changes in order amounts while accounting for missing data.



Count: **5630**	Min: **11**
Distinct: **17**	Max: **26**
Duplicate: **5613**	Mean: **15.71**
Valid: **5365**	Variance: **13.51**
Empty: **265**	Median: **15**
Invalid: **0**	Lower quantile: **13**
	Upper quantile: **18**

3.1.14 CouponUsed

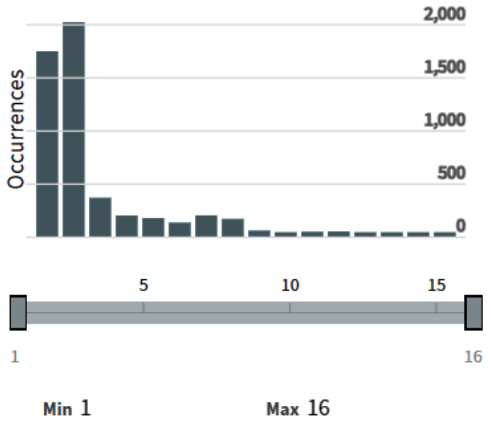
CouponUsed captures the count of coupons utilized by customers in the last month. Analyzing coupon usage patterns might reveal correlations between promotional offers and customer retention or churn rates. Similar to OrderAmountHikeFromlastYear, this attribute had 265 missing values addressed through mean imputation using the mean value of 2 derived from the exploration phase.



Count: **5630**	Min: **0**
Distinct: **18**	Max: **16**
Duplicate: **5612**	Mean: **1.75**
Valid: **5374**	Variance: **3.59**
Empty: **256**	Median: **1**
Invalid: **0**	Lower quantile: **1**
	Upper quantile: **2**

3.1.15 OrderCount

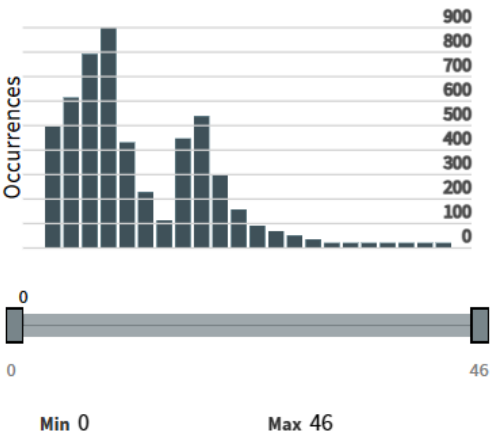
OrderCount represents the total count of orders placed by customers in the last month. Analysis of order frequency might highlight its role in predicting churn and its influence on customer engagement and loyalty. Similar to CouponUsed and other attributes, this attribute experienced 258 missing values addressed through mean imputation using the mean value of 3 derived during the exploration phase.



Count: **5630**	Min: **1**
Distinct: **17**	Max: **16**
Duplicate: **5613**	Mean: **3.01**
Valid: **5372**	Variance: **8.64**
Empty: **258**	Median: **2**
Invalid: **0**	Lower quantile: **1**
	Upper quantile: **3**

3.1.16 DaySinceLastOrder

DaySinceLastOrder indicates the duration since the customer's last order. Understanding this attribute's relationship with churn could elucidate the importance of regular customer activity in retention strategies. Similar to Tenure, HourSpendOnApp, CouponUsed, OrderCount, and OrderAmountHikeFromlastYear, this attribute encountered 307 missing entries addressed through mean imputation using the mean value of 5 derived from initial exploration. This approach maintains insights into the recency of customer transactions within the platform.



Count: **5630**	Min: **0**
Distinct: **23**	Max: **46**
Duplicate: **5607**	Mean: **4.54**
Valid: **5323**	Variance: **13.35**
Empty: **307**	Median: **3**
Invalid: **0**	Lower quantile: **2**
	Upper quantile: **7**

3.1.18 Attributes Excluded:

Certain attributes, including WarehouseToHome, NumberOfDeviceRegistered, NumberOfAddress, and CashbackAmount were deemed extraneous to the churn prediction analysis due to their limited relevance. These attributes lacked substantive impact or failed to provide actionable insights essential for predicting customer churn behavior within the e-commerce platform.

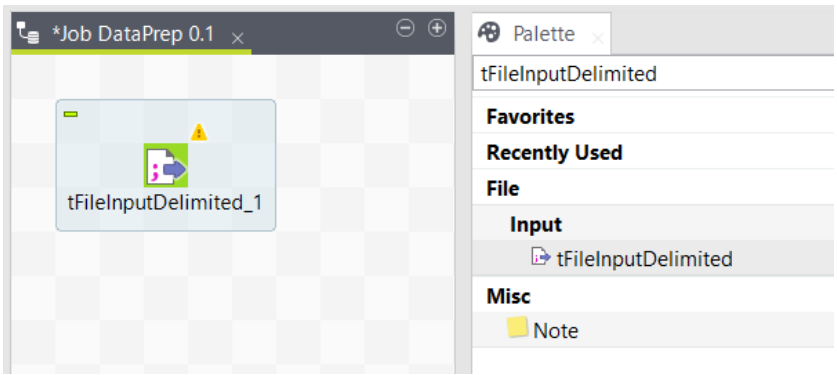
3.2 Data Preprocessing & Transformation

The Data Preprocessing & Transformation phase was meticulously executed using Talend Data Integration, employing various techniques to ensure data quality, handle missing values, and streamline the dataset for effective predictive modeling.

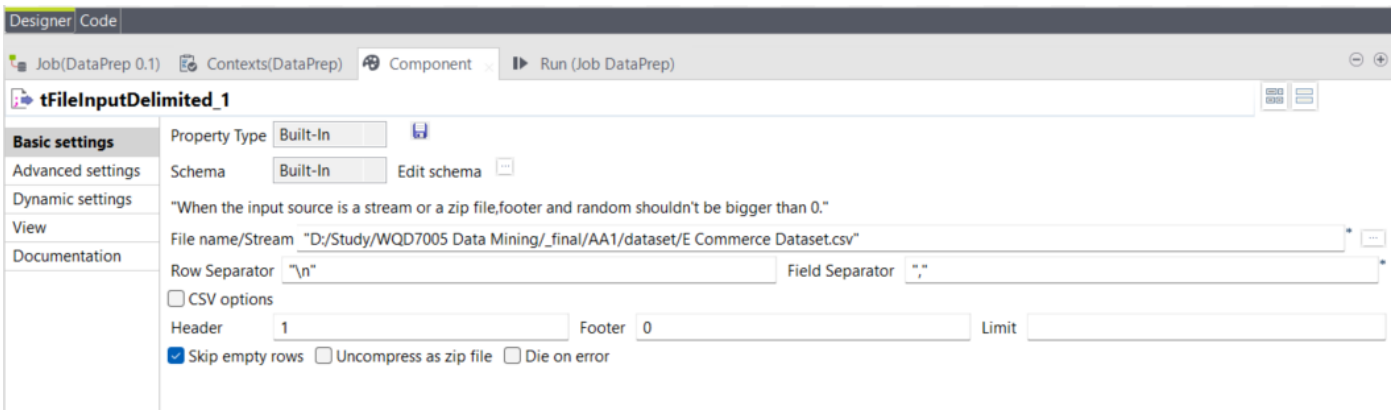
3.3.1 Import Data

In order to import the raw dataset into Talend Data Integration:

- First, we will find and drag the component "tFileInputDelimited" from the Palette into the design workspace



- Then, we will configure the component by changing the Field Separator to comma "," and specify the Header at row 1



- Lastly, we will click on the "Edit schema" button to add columns according to the features of our raw dataset discovered during the Data Exploration phase

Schema of tFileInputDelimited_1

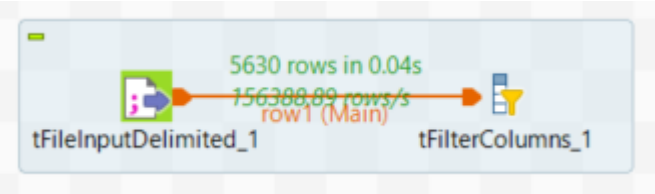
tFileInputDelimited_1

Column	Key	Type	N..	Date Pattern (Ctrl+Sp...	Length	Precision	Default	Comment
CustomerID	☑	int	☐					
Churn	☐	boolean	☐					
Tenure	☐	Integer	☑					
PreferredLoginDevice	☐	String	☑					
CityTier	☐	Integer	☑					
WarehouseToHome	☐	Integer	☑					
PreferredPaymentMode	☐	String	☑					
Gender	☐	String	☑					
HourSpendOnApp	☐	Integer	☑					
NumberOfDeviceRegistered	☐	Integer	☑					
PreferedOrderCat	☐	String	☑					
SatisfactionScore	☐	Integer	☑					
MaritalStatus	☐	String	☑					
NumberOfAddress	☐	Integer	☑					
Complain	☐	Boolean	☑					
OrderAmountHikeFromlastYear	☐	Integer	☑					
CouponUsed	☐	Integer	☑					
OrderCount	☐	Integer	☑					
DaySinceLastOrder	☐	Integer	☑					
CashbackAmount	☐	Integer	☑					

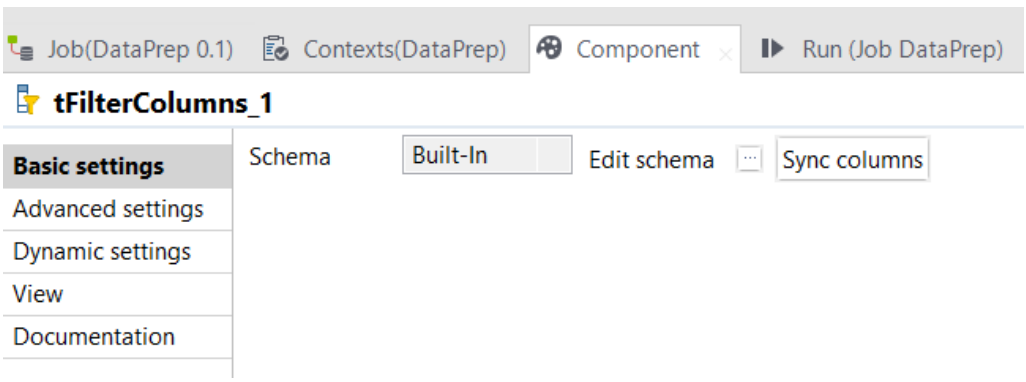
3.2.2 Dropping Irrelevant Attributes

Attributes that were deemed inconsequential to predicting customer churn were systematically eliminated from the dataset to streamline further analysis. Attributes such as WarehouseToHome, NumberOfDeviceRegistered, NumberOfAddress and CashbackAmount lacked substantial influence or relevance in discerning churn behavior. Their exclusion aimed to refine the dataset, focusing solely on attributes with the potential to significantly impact churn prediction.

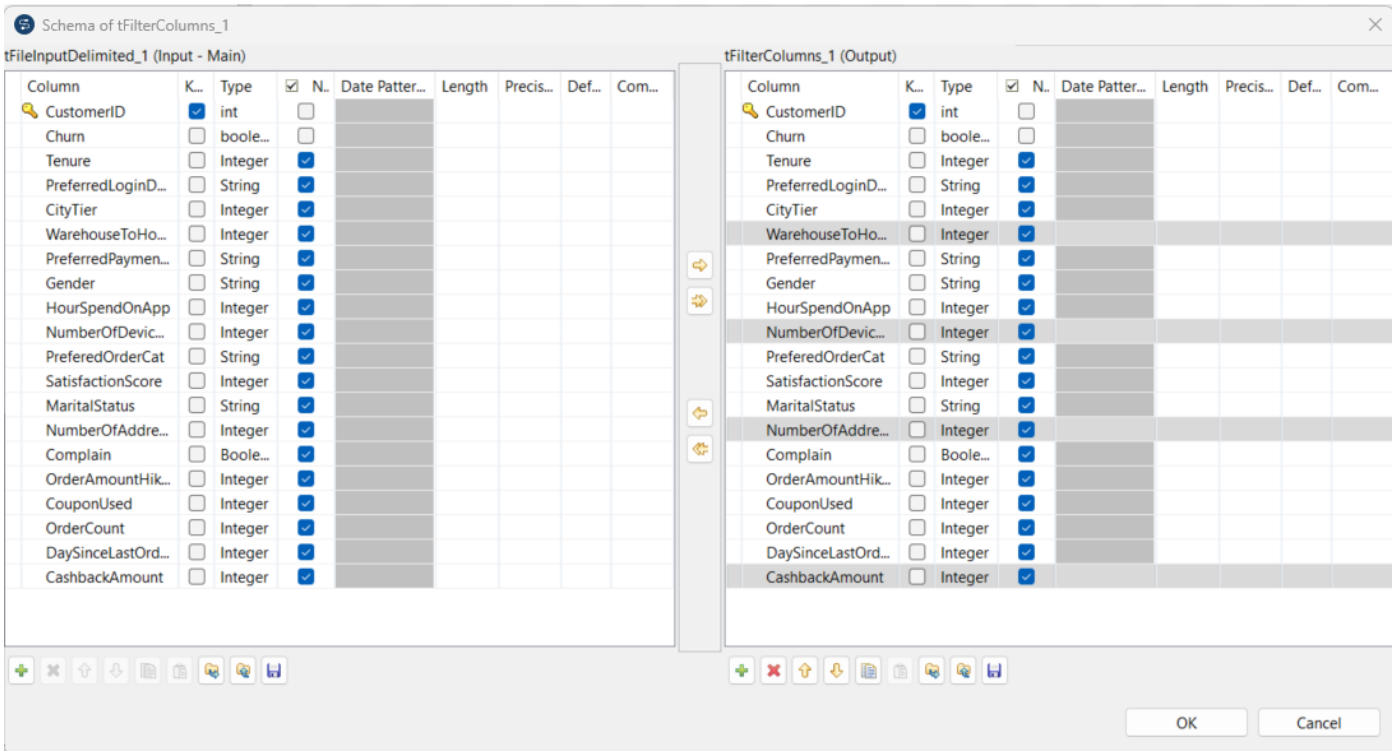
- First, we will find and drag the component "tFilterColumns" from the Palette into the design workspace, and connect "tFilterColumns" to "tFileInputDelimited"



- Then, we will configure the component. Click the "Sync columns" button to include all columns in our component output, and click the "Edit schema" button to view the schema



- At the Schema of the tFilterColumns component, use Ctrl+Click to select the columns that we wish to drop at the output



- Lastly, click the "OK" button to confirm the settings

3.2.2.1 Dropping: WarehouseToHome

The WarehouseToHome attribute, representing the distance between the warehouse and customers' homes, was omitted due to ambiguity regarding the unit of measurement. The dataset lacked clarification regarding distance units (e.g., miles, kilometers), rendering this attribute inconclusive for meaningful churn analysis. Its exclusion prevented potential misinterpretation or incorrect assumptions based on unclear distance units, ensuring the dataset's reliability.

3.2.2.2 Dropping: NumberOfDeviceRegistered

The NumberOfDeviceRegistered attribute, denoting the total number of devices registered by a customer, was deemed irrelevant for predicting churn behavior. This attribute lacked substantial relevance or evident correlation with customer churn. Its exclusion aimed to streamline the dataset, eliminating factors with minimal or no influence on predicting customer attrition, thus optimizing the modeling process.

3.2.2.3 Dropping: NumberOfAddress

Similar to NumberOfDeviceRegistered, the NumberOfAddress attribute, depicting the total number of addresses added by a customer, was considered irrelevant for predicting churn within the e-commerce platform. Its lack of substantive impact or correlation with customer behavior made it inconsequential for discerning churn patterns. Removing this attribute focused the analysis on factors with more pronounced influence on predicting churn behavior.

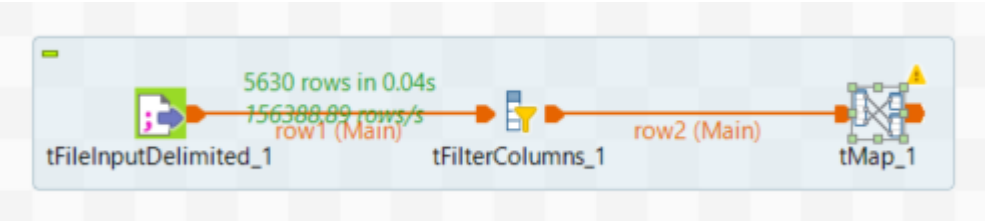
3.2.2.4 Dropping: CashbackAmount

The CashbackAmount attribute, representing the average cashback received by customers, was omitted due to the absence of specified currency units. Without a defined currency, the attribute's numerical values lacked contextual relevance for meaningful analysis. Its exclusion prevented potential misinterpretation or skewed analysis due to unspecified currency units, ensuring data accuracy and reliability in the churn prediction process.

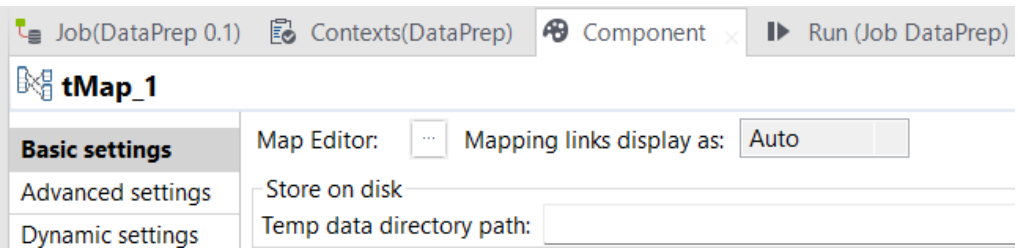
3.2.3 Handling Missing Values

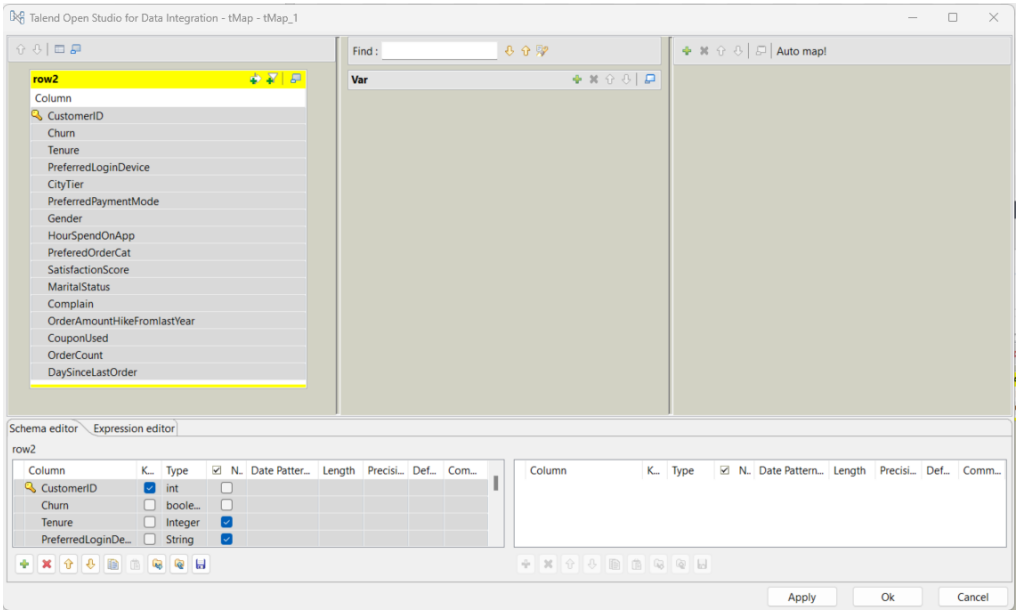
When addressing missing values in the dataset, a mean imputation strategy was implemented for columns with continuous attributes. The decision to use mean imputation was based on the nature of the missing data and the continuous nature of these specific attributes, ensuring a method that maintained the dataset's integrity while filling in the missing information.

- First, we will find and drag the component "tMap" from the Palette into the design workspace, and connect "tMap" to "tFilterColumns"

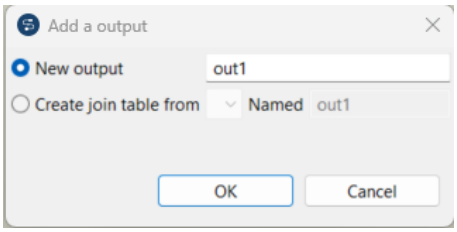


- Then, click on the "Map Editor" button to configure the tMap component

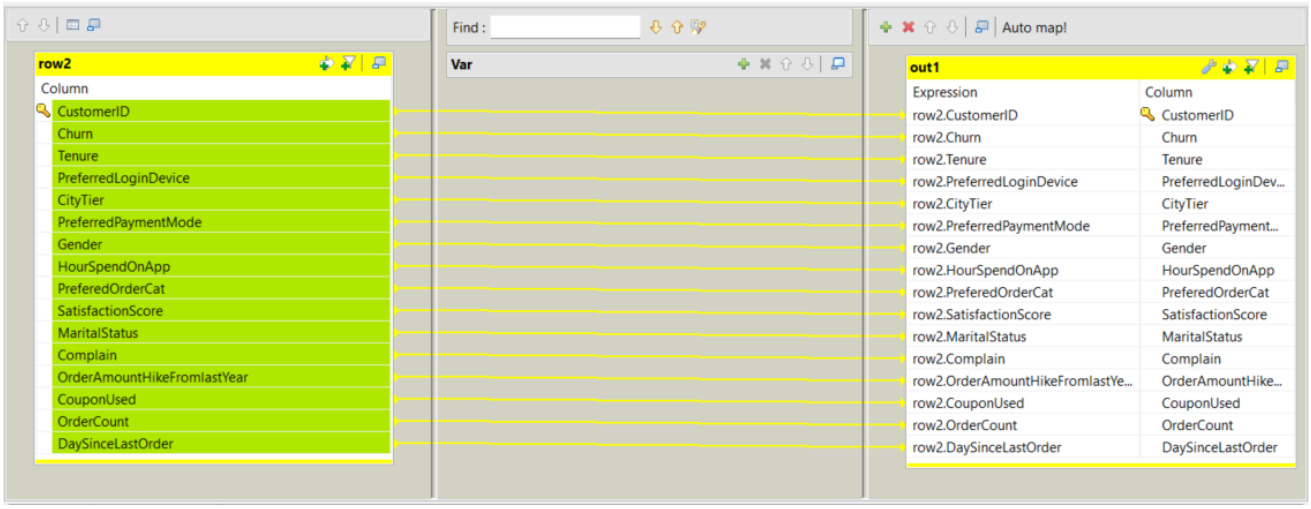




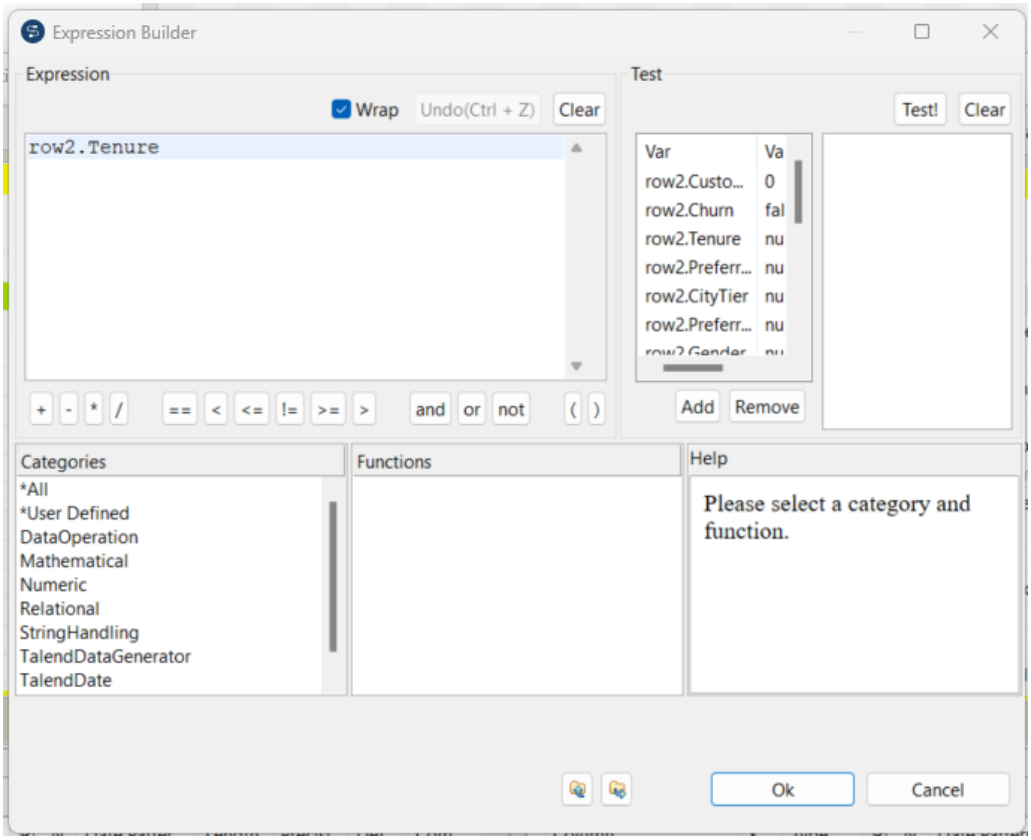
- Create on the "+" button on the third section of the editor to add a new output



- Drag all the columns from the input table "row1" to the output table "out1"

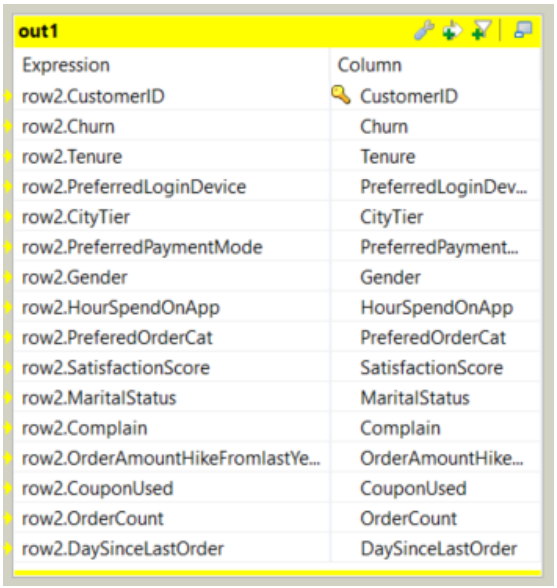
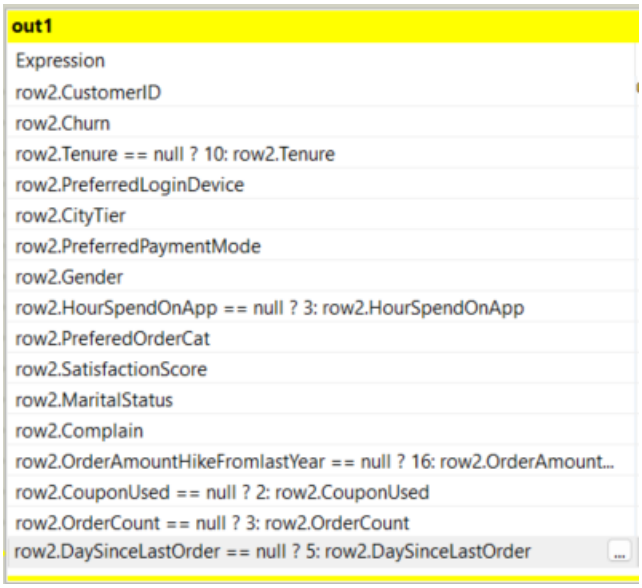


- Click on the column of the output table to open the "Expression Builder"



- Apply the following expression to all the columns that we need to handle the missing values

```
row2.Tenure == null ? 10: row2.Tenure
```

Before	After
	

3.2.3.1 Imputing: Tenure

To address the 264 missing values within the Tenure attribute, a mean imputation technique was utilized. The mean value of 10, derived during the data exploration phase, was employed to fill the missing entries. This approach aimed to ensure that the imputed values were

representative of the typical association duration, maintaining the attribute's integrity for subsequent analysis.

3.2.3.2 Imputing: HourSpendOnApp

Similarly, for the HourSpendOnApp attribute, which encountered 255 missing values, mean imputation using the calculated mean value of 3 was implemented. This process ensured a comprehensive dataset without compromising insights into customers' engagement levels on the mobile application or website.

3.2.3.3 Imputing: OrderAmountHikeFromlastYear

Addressing the 265 missing entries in the OrderAmountHikeFromlastYear attribute, mean imputation based on the mean value of 16 was applied. This technique aimed to maintain the attribute's representation of percentage increases in order amounts from the previous year while accounting for missing data.

3.2.3.4 Imputing: CouponUsed

Imputation of missing values (265 instances) in the CouponUsed attribute was conducted by utilizing the mean value of 2 derived from the exploration phase. This process ensured that the imputed values reflected the typical count of coupons used in the absence of specific data.

3.2.3.5 Imputing: OrderCount

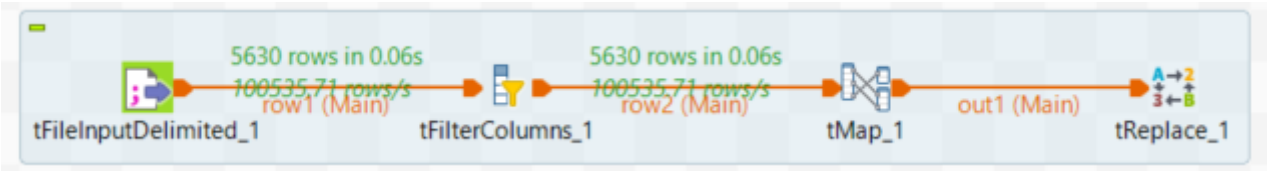
For the OrderCount attribute, encompassing 258 missing values, mean imputation utilizing the mean value of 3 was employed. This method ensured a comprehensive dataset, maintaining insights into the frequency of orders placed by customers.

3.2.3.6 Imputing: DaySinceLastOrder

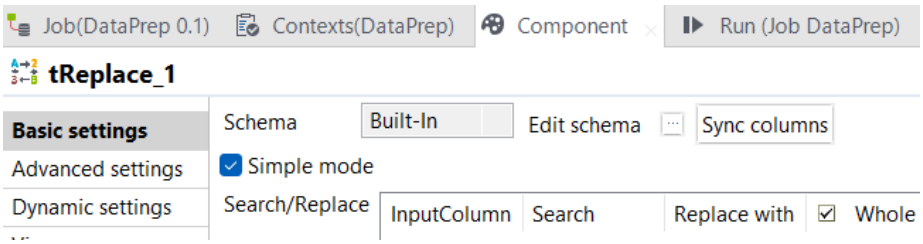
The DaySinceLastOrder attribute, which encountered 307 missing entries, underwent mean imputation using the mean value of 5 derived from initial exploration. This approach aimed to retain insights into the recency of customer transactions within the platform.

3.2.4 Transformation of Categorical Variables

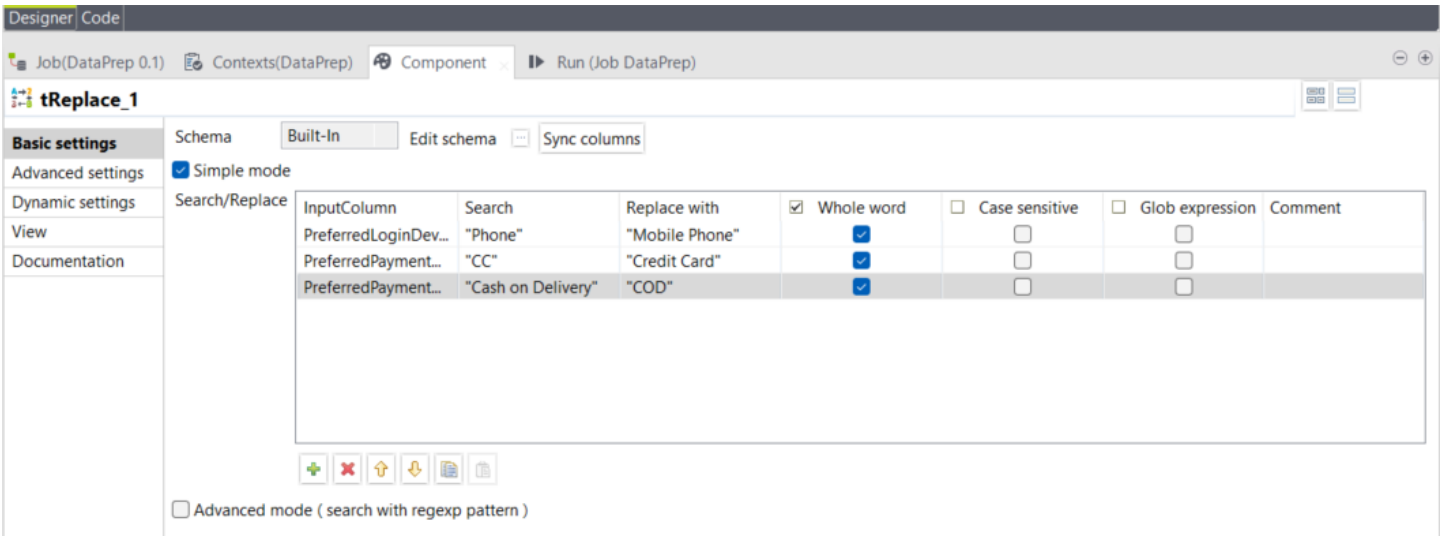
- First, we will find and drag the component "tReplace" from the Palette into the design workspace, and connect "tReplace" to "tMap"



- Then, we will configure the component. Click the "Sync columns" button to include all columns in our component output



- Lastly, add the columns to perform the replacement, the targeted word (Search) and the replacement (Replace with)



3.2.4.1 Transform: PreferredLoginDevice

The PreferredLoginDevice attribute delineated customers' preferences for logging into the e-commerce platform. Recognizing the need for streamlined categories to enhance classification accuracy, a specific transformation was applied. The values 'Phone' and 'Mobile Phone,' denoting similar contexts, were aggregated into a unified category, 'Mobile Phone.' This consolidation aimed to harmonize and simplify device preferences, facilitating clearer distinctions and more accurate predictions within the modeling process.

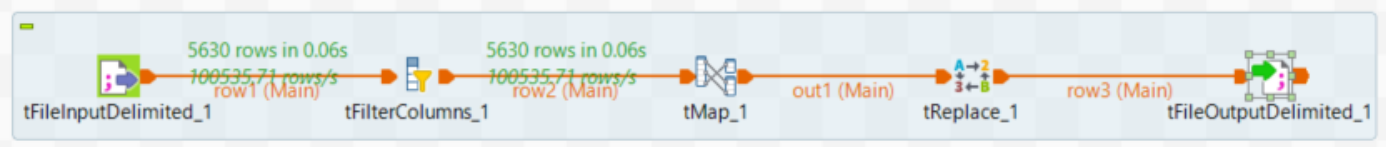
3.2.4.3 Transform: PreferredPaymentMode

The PreferredPaymentMode attribute delineated various payment modes favored by customers. To standardize and refine the payment mode categories for improved model interpretability, two transformations were implemented. The abbreviation 'CC' was unified into 'Credit Card,' ensuring consistency within the category. Similarly, 'Cash on Delivery' was transformed into 'COD,' aligning payment mode descriptions to streamline the dataset. These transformations aimed to standardize payment preferences, reducing ambiguity and enhancing the attribute's efficacy in predicting churn behavior.

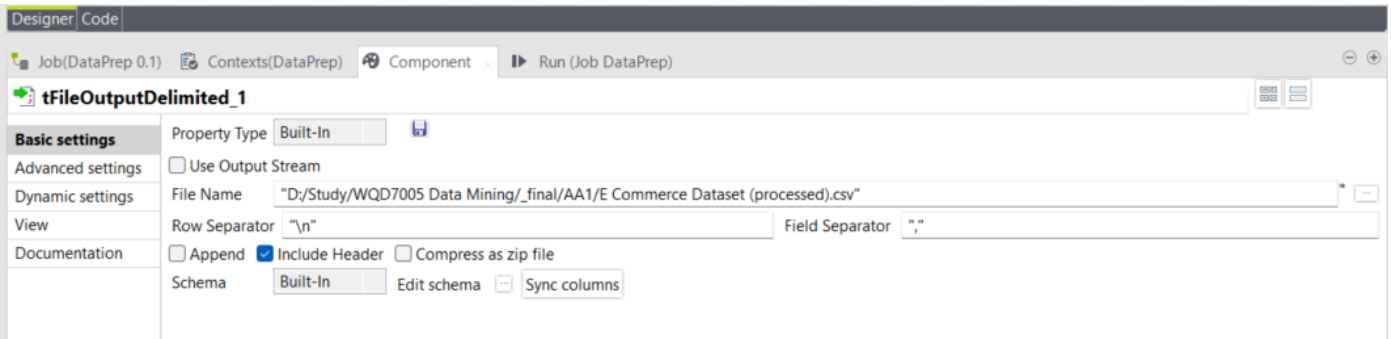
3.2.5 Export Data

The final step involves exporting the preprocessed dataset for subsequent modeling and analysis.

- First, we will find and drag the component "tFileOutputDelimited" from the Palette into the design workspace, and connect "tFileOutputDelimited" to "tReplace"



- Then, we will configure the component by changing the Field Separator to comma "," and tick the Include Header checkbox



- Lastly, we will run the entire job to export our cleaned dataset

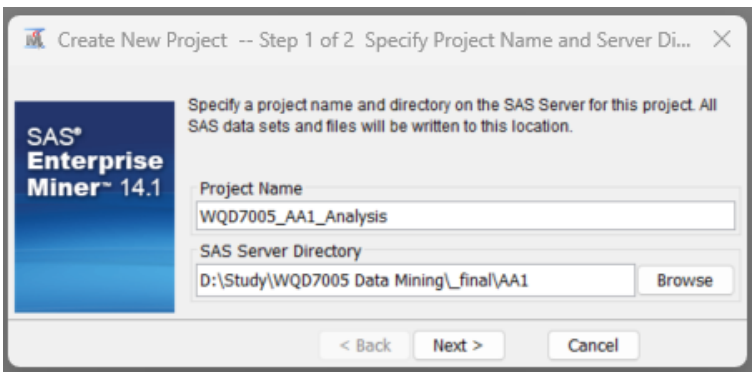
3.3 Data Analysis

SAS Enterprise Miner served as the primary tool for this phase. Its robust capabilities facilitated the creation of decision trees, ensemble methods such as Random Forest and Gradient Boost, and comprehensive model comparisons. This platform enabled a streamlined approach to predictive modeling, empowering the exploration of customer churn behavior within the e-commerce domain with precision and efficiency.

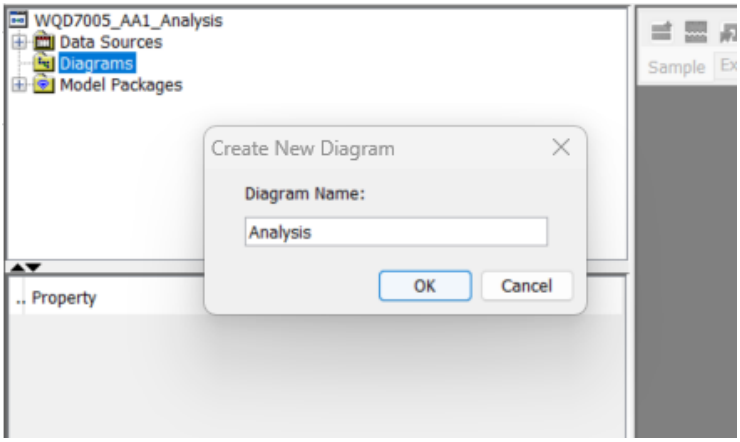
3.3.1 Import Data

The initial step in the Data Analysis involved importing the processed dataset into the analysis platform:

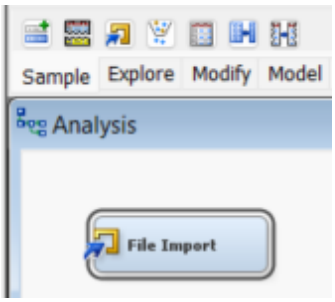
- First, we will create a new SAS project



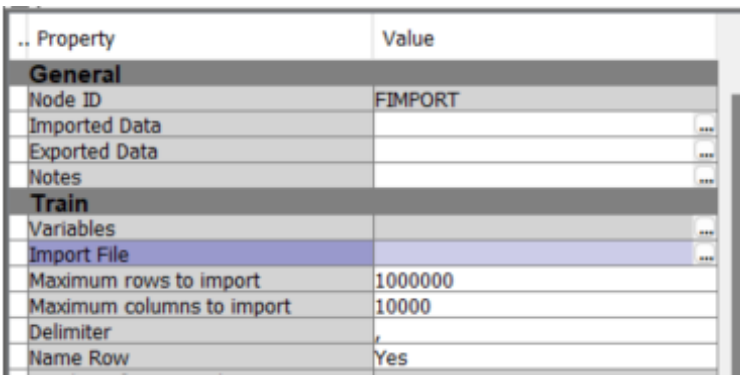
- Next, we will create a new diagram for our SAS project

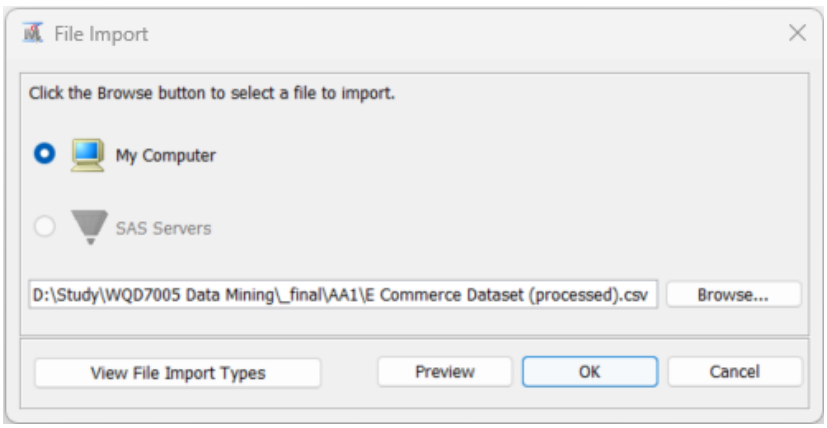


- From the Sample palette, drag the "File Import" node into the diagram workspace

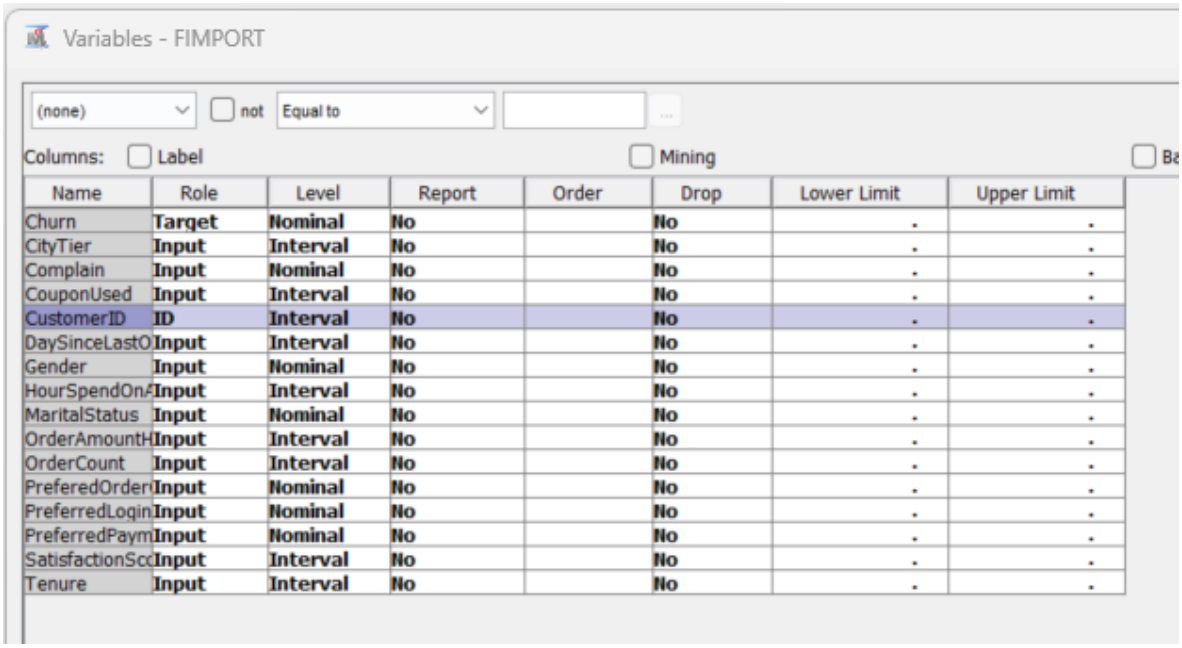


- Double-click on the "File Import" node to open its properties. Click on the "Import File" property and choose the processed dataset file





- Right click the "File Import" node, and select the "Edit Variables" option. Then, set Churn as Target and CustomerID as ID

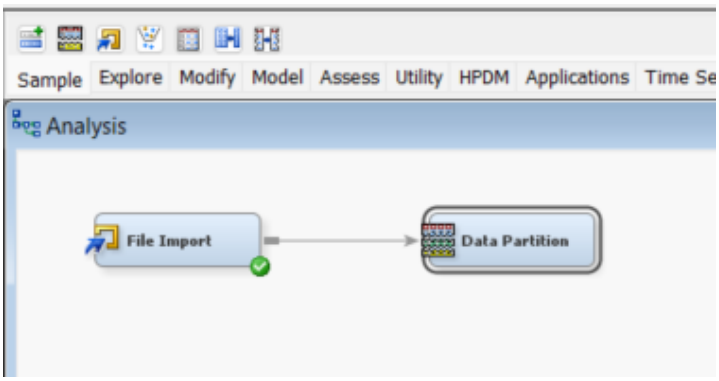


Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Churn	Target	Nominal	No		No	.	.
CityTier	Input	Interval	No		No	.	.
Complain	Input	Nominal	No		No	.	.
CouponUsed	Input	Interval	No		No	.	.
CustomerID	ID	Interval	No		No	.	.
DaySinceLastO	Input	Interval	No		No	.	.
Gender	Input	Nominal	No		No	.	.
HourSpendOnA	Input	Interval	No		No	.	.
MaritalStatus	Input	Nominal	No		No	.	.
OrderAmountH	Input	Interval	No		No	.	.
OrderCount	Input	Interval	No		No	.	.
PreferedOrder	Input	Nominal	No		No	.	.
PreferredLogin	Input	Nominal	No		No	.	.
PreferredPaym	Input	Nominal	No		No	.	.
SatisfactionSc	Input	Interval	No		No	.	.
Tenure	Input	Interval	No		No	.	.

3.3.2 Train-Test Splitting

The next step is followed by data partitioning. The dataset underwent division into subsets for training, validation, and testing. This partitioning, with 70% for training and 30% for validation, aimed to ensure a robust model evaluation process and minimize overfitting.

- From the Sample palette, drag the "Data Partition" node into the workspace, and connect it to the "File Import" node



- Double-click on the "Data Partition" node to open its properties. Set the "Training" property to 70.0, "Validation" property to 30.0, and "Test" property to 0.0

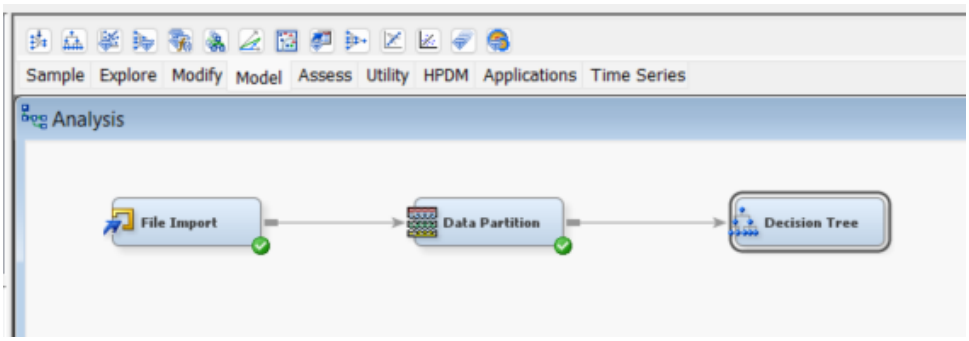
Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	1/7/24 1:37 AM
Run ID	

3.3.3 Modeling

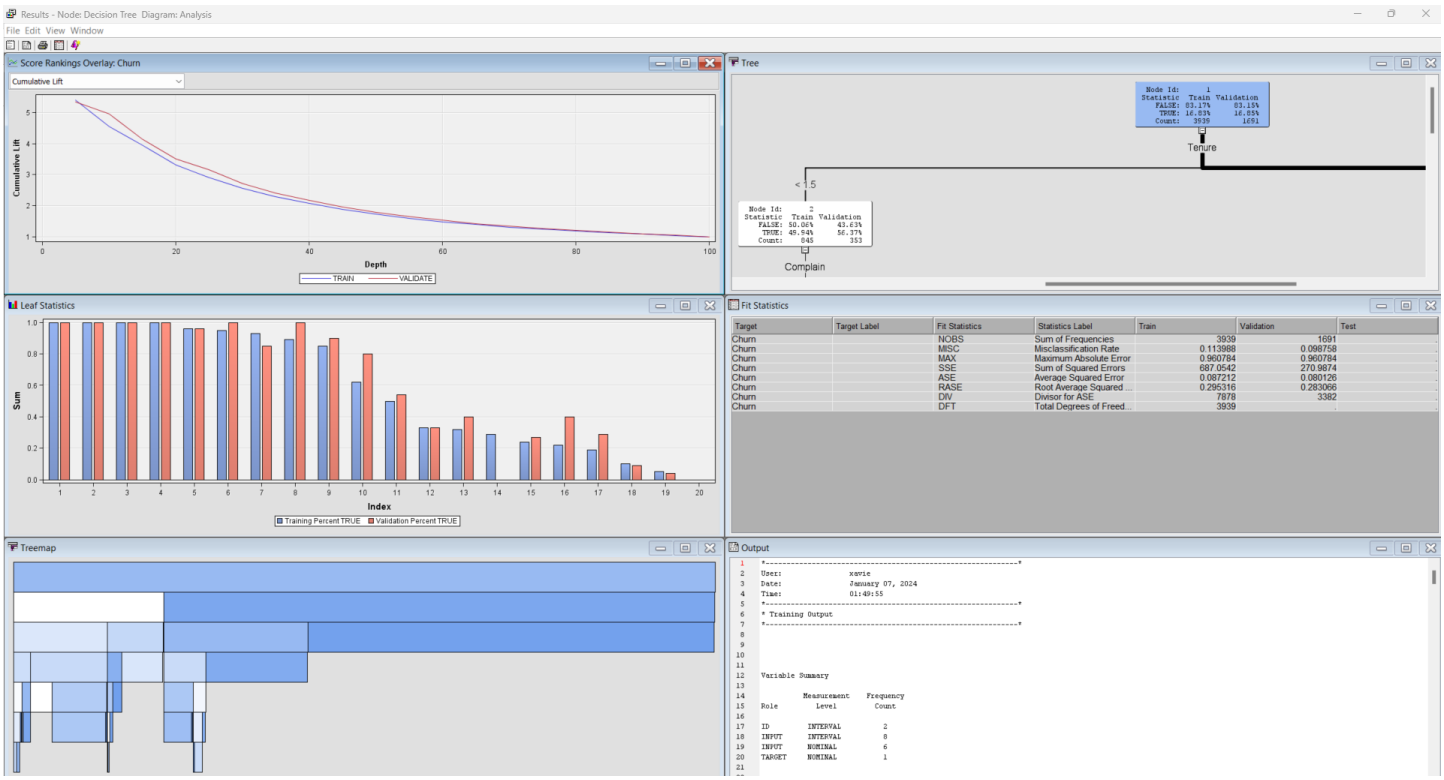
3.3.3.1 Decision Tree

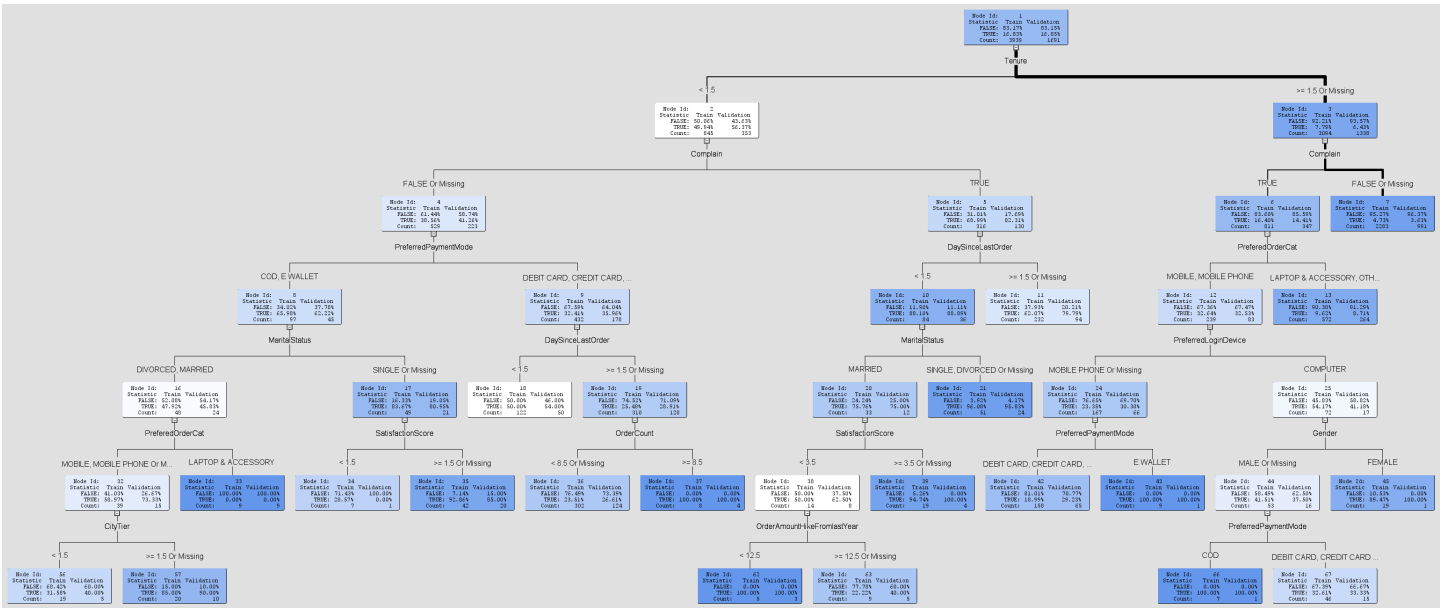
The modeling phase commenced with the construction of a decision tree model using the processed dataset. This decision tree aimed to map out potential decision pathways based on attribute conditions, enabling the understanding of significant predictors influencing customer churn within the e-commerce platform.

- From the Model palette, drag the "Decision Tree" node into the workspace, and connect it to the "Data Partition" node



- Run the "Decision Tree" node with its default properties, and view the results

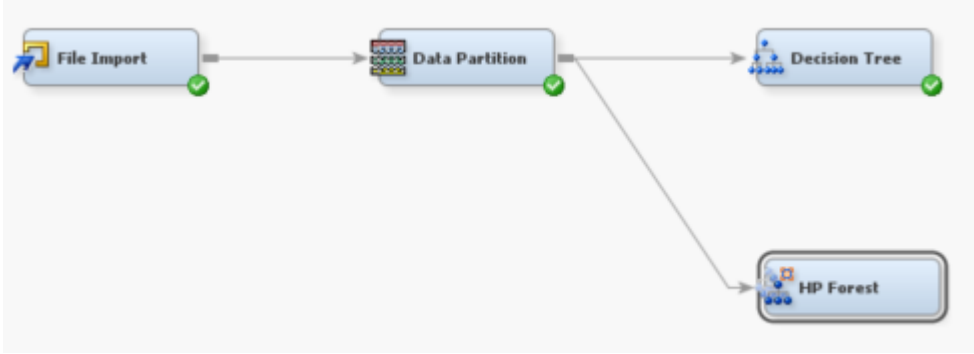




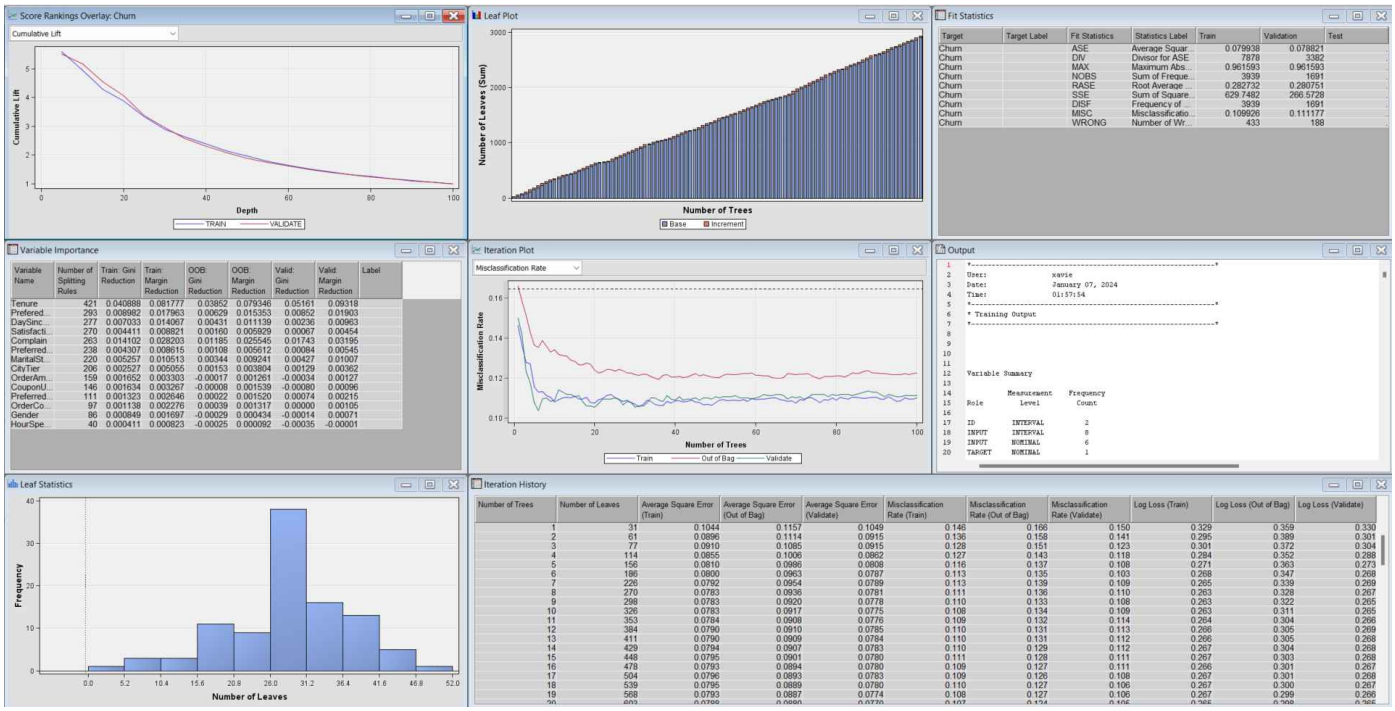
3.3.3.2 Bagging: Random Forest

Utilizing bagging techniques, specifically the Random Forest algorithm, an ensemble of decision trees was constructed. This method aggregated multiple decision trees to create a more robust and accurate predictive model, leveraging the diversity of individual trees to improve overall predictive power.

- From the HDPM palette, drag the "HP Forest" node into the workspace, and connect it to the "Data Partition" node



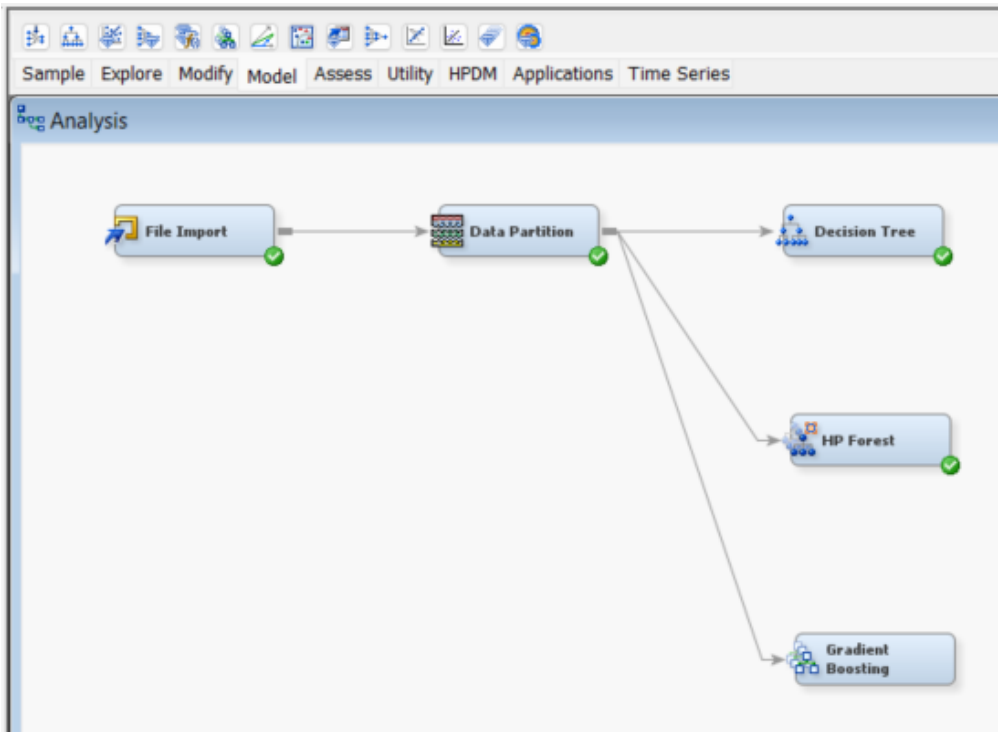
- Run the "HP Forest" node with its default properties, and view the results



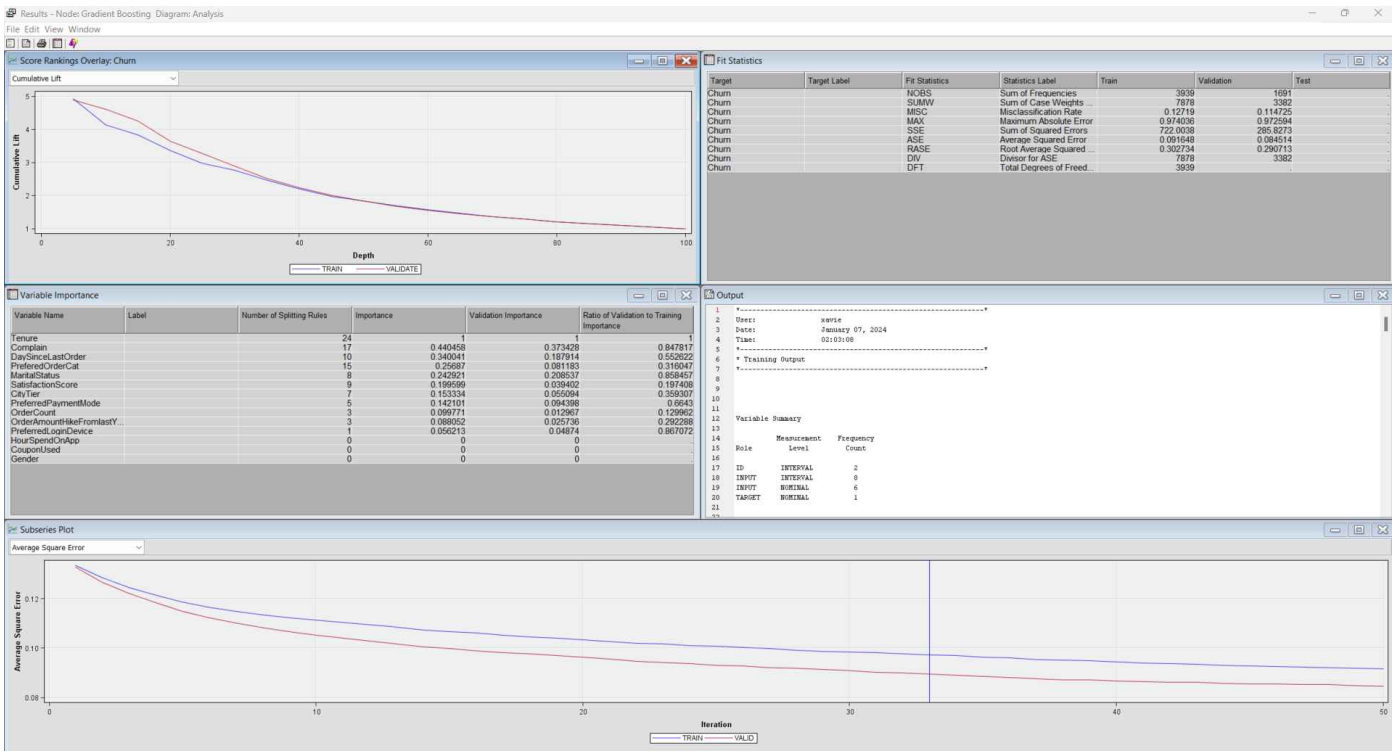
3.3.3.3 Boosting: Gradient Boost

Incorporating boosting methodologies, specifically the Gradient Boosting algorithm, another ensemble model was developed. Gradient Boosting sequentially constructed decision trees, each addressing the weaknesses of its predecessor, ultimately enhancing the predictive accuracy by focusing on previously misclassified instances.

- From the Model palette, drag the "Gradient Boosting" node into the workspace, and connect it to the "Data Partition" node



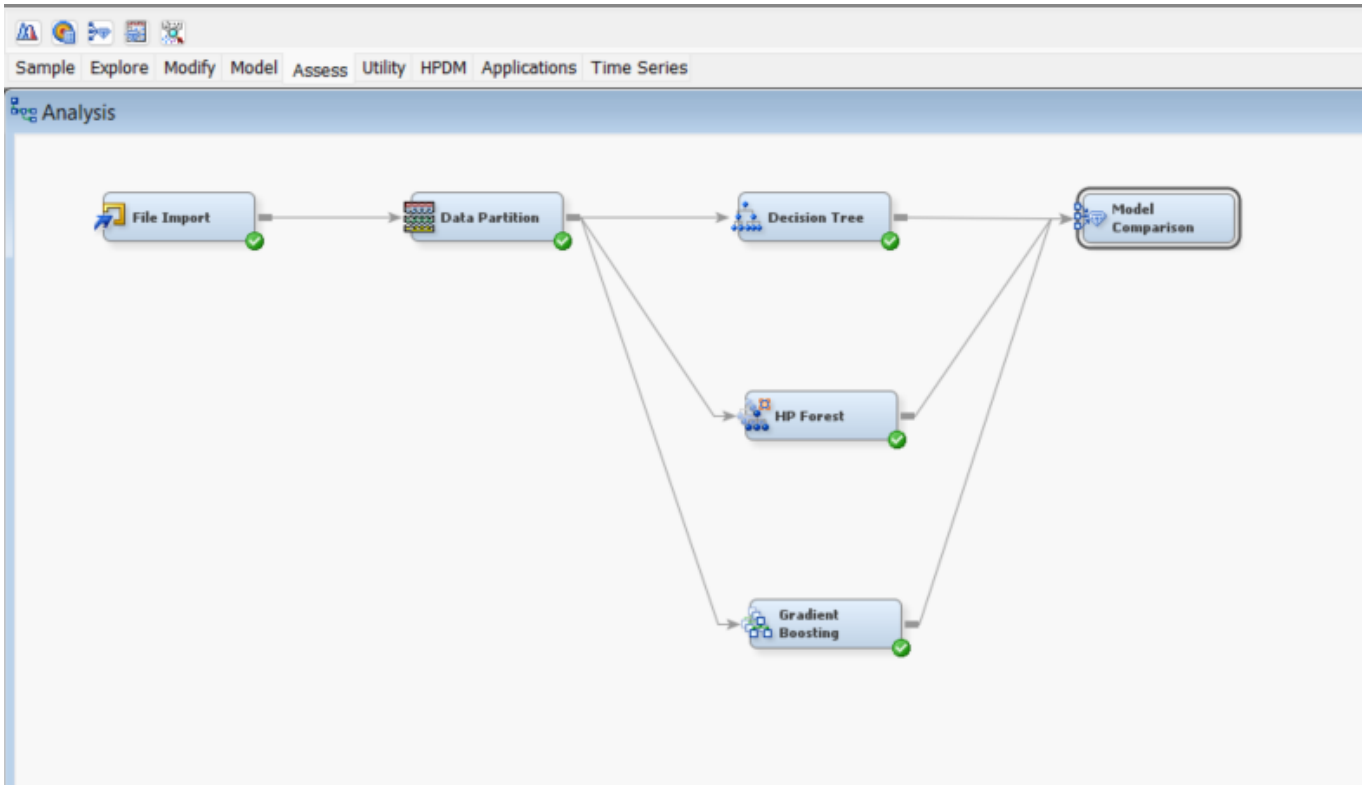
- Run the "Gradient Boosting" node with its default properties, and view the results



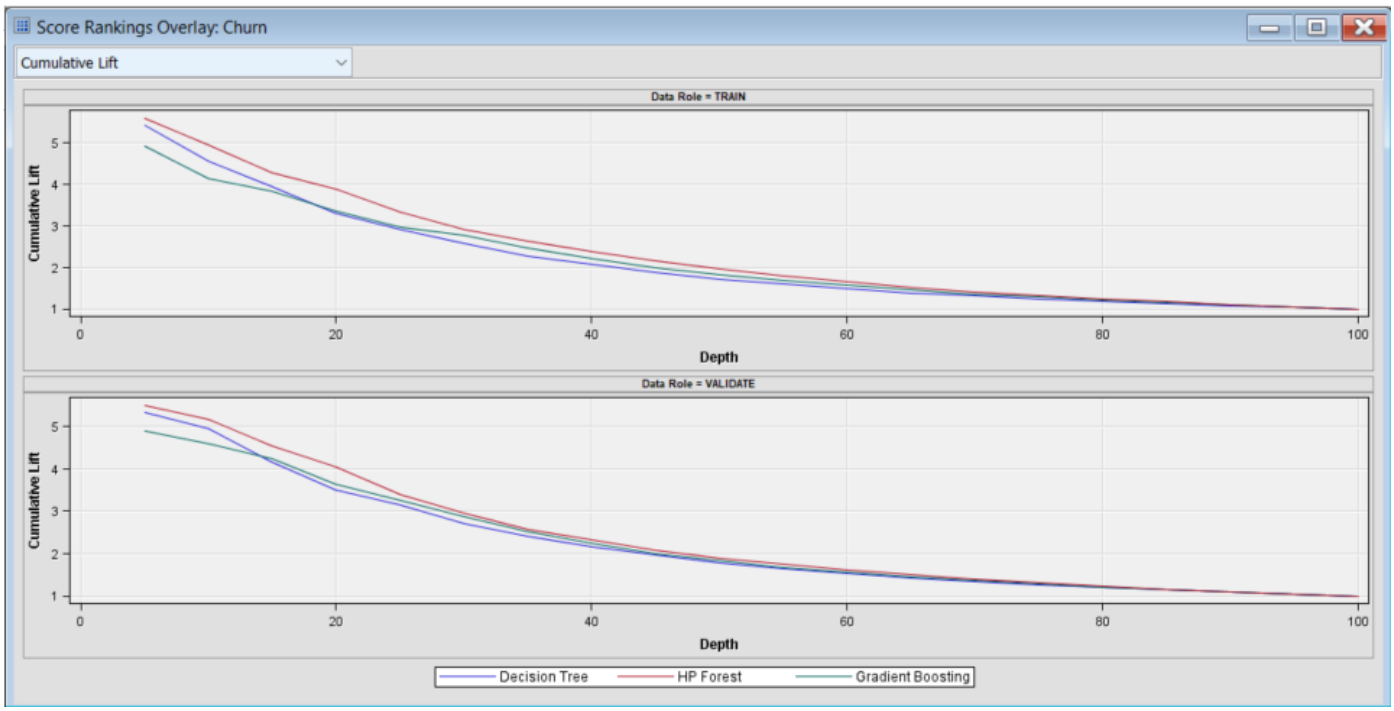
3.3.3 Evaluation

The evaluation stage involved a comprehensive comparison of the developed models. Various performance metrics, including misclassification rates, squared errors, and other relevant evaluation criteria, were employed to compare and contrast the effectiveness of the decision tree, Random Forest (bagging), and Gradient Boost models. This comparative analysis aimed to identify the most accurate and robust model for predicting customer churn within the e-commerce platform.

- From the Assess palette, drag the "Model Comparison" node into the workspace, and connect it to the all the model nodes



- Run the node and view the results



Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree	Decision Tree	0.09876	0.087212	0.11399	0.080126
	HPDMForest	HP Forest	0.11118	0.079938	0.10993	0.078821
	Boost	Gradient Boosting	0.11473	0.091648	0.12719	0.084514

Chapter 4: Result & Discussion

4.1 Model Comparison

The evaluation of multiple models—Decision Tree, Random Forest (bagging), and Gradient Boost—revealed critical insights into their performance metrics. The Decision Tree emerged as the most accurate and robust model for predicting customer churn within the e-commerce platform. Its superior interpretability and accuracy surpassed the ensemble methods of Random Forest and Gradient Boosting, signifying its efficacy in predicting customer attrition with precision and reliability.

4.2 Feature Importance

In the Decision Tree model, attribute importance analysis illuminated the crucial predictors influencing customer churn. Key attributes such as Tenure, Complain, PreferredPaymentMode, PreferedOrderCat, DaySinceLastOrder, and PreferredLoginDevice exhibited substantial influence in predicting churn behavior. These attributes play pivotal roles in deciphering customer retention strategies and shaping targeted marketing approaches, underlining their significance in steering effective business decisions.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Tenure		1	1.0000	1.0000	1.0000
Complain		2	0.4737	0.4823	1.0181
PreferredPaymentMode		3	0.3827	0.2074	0.5419
PreferedOrderCat		2	0.3120	0.3074	0.9855
DaySinceLastOrder		2	0.2829	0.1823	0.6443
PreferredLoginDevice		1	0.2013	0.0000	0.0000
OrderCount		1	0.1966	0.1765	0.8976
SatisfactionScore		2	0.1863	0.1042	0.5594
MaritalStatus		2	0.1825	0.1613	0.8837
Gender		1	0.1652	0.0953	0.5771
CityTier		1	0.1536	0.1262	0.8217
OrderAmountHikeFromlastYear		1	0.1284	0.0972	0.7569

4.3 Discussion and Implications

The results underscore the significance of specific attributes in predicting customer churn within the e-commerce landscape. These insights pave the way for strategic implications aimed at enhancing customer retention strategies, refining targeted marketing approaches, and improving service quality. The identified influential attributes serve as focal points for developing proactive business strategies geared toward bolstering customer satisfaction, retention, and overall business growth in the competitive e-commerce domain.

Chapter 5: Conclusion

The analysis embarked on a comprehensive exploration of e-commerce customer behavior, employing advanced data mining techniques to predict and understand customer churn within the platform. The critical insights derived from this study offer substantial value to augment business strategies and enhance customer retention efforts within the dynamic e-commerce landscape.

5.1 Summary of Key Findings

The analysis illuminated pivotal aspects:

- The Decision Tree model emerged as the most accurate in predicting customer churn, outperforming ensemble methods like Random Forest and Gradient Boost.
- Attributes such as Tenure, Complain, PreferredPaymentMode, PreferedOrderCat, DaySinceLastOrder, and PreferredLoginDevice were identified as crucial predictors significantly influencing customer churn.

5.2 Implications for Business Strategy

The identified influential attributes lay the foundation for strategic business implications:

- Enhancing Customer Retention Strategies: Focus on improving services based on insights derived from attributes like Tenure and PreferredPaymentMode.
- Targeted Marketing Approaches: Tailoring marketing strategies using PreferedOrderCat and PreferredLoginDevice to cater to specific customer preferences.
- Service Quality Enhancements: Addressing customer concerns raised through Complain and DaySinceLastOrder metrics to enhance overall service quality.

Appendix

GitHub link: https://github.com/jvloo/WQD7005_AA1/