Module Code: CT045-3-M-ABAV

Module Name: Advanced Business Analytics & Visualization

Individual Assignment

Problem Analysis & Solution Development

Student Name: Muhammad Yousouf Ali Budullah
TP Number: TP086704
Programme: MSc Data Science & Business Analytics
Intake Code: APDMF2501DSBA(BI)(PR)
Module Lecturer: Dr. Preethi Subramanian

Date of Submission: 16 September 2025

Contents

1 Problem Analysis

1.1 Business Context

Banking today is a very competitive industry , where customers have a number of options available to them which include online banking and fintech services. As a result, customer churn, when clients leave the bank, has become a serious problem as it happens frequently and directly impacts a banks profit. Every customer that leaves the bank means less revenue and much larger costs to replace them. According to (de Lima Lemos et al., 2022) acquiring new customers can cost up to five times more than keeping existing ones, which proves churn is not only common but also expensive. This makes churn a critical problem for banks to address, as losing customers prevents cross selling opportunities and long-term customer value. Therefore, a data-driven approach is required to predict said churn and prevent it before it occurs. It is also needed to understand why customer are leaving, so that they can be proactive and more effective with their retention strategies.

1.2 Dataset and Sources

The Bank Customer Churn dataset is used in this project, containing 10,000 customer records with 13 features that describe demographics, account details, and the churn result. It is provided through the Maven Analytics Data Playground and is also hosted on Kaggle, which Maven cites as its source. The dataset is publicly available and used for machine learning and modeling tasks for churn prediction.

Feature	**Description**	**Type**
CustomerID	Unique identifier for each customer	Identifier
Surname	Customer last name	Categorical
CreditScore	Numerical value representing credit rating	Numeric
Geography	Customer location (France, Spain, Germany)	Categorical
Gender	Customer gender (Male/Female)	Categorical
Age	Age of the customer	Numeric
Tenure	Years the customer has been with the bank	Numeric
Balance	Current account balance	Numeric
NumOfProducts	Number of bank products used (1–4)	Numeric
HasCrCard	Credit card ownership (1 = Yes, 0 = No)	Binary
IsActiveMember	Active membership (1 = Yes, 0 = No)	Binary
EstimatedSalary	Estimated annual salary	Numeric
Exited	Target: 1 = churned, 0 = stayed	Binary

Table 1: Features in the Bank Customer Churn dataset.

Note. Dataset obtained from Maven Analytics Data Playground and Kaggle.

1.3 Business Problems and Objectives

The main business problem that banks are facing is that they are losing customers without a reliable way to identify the risk of churn in advance. Retention campaigns used are often vague and do not target the customers susceptible to churn. This not only wastes resources it also fails to prevent valuable customers from leaving. Without analytics, churn management is mostly reactive and not proactive, causing banks to miss the right timing to retain their customers. This project aims to use SAS Viya to create a pipeline in order to predict churn, select the best performing model, and interpret the key drivers of churn, so banks can take informed and targeted decisions to improve customer retention.

Objectives

1. Develop predictive models in SAS Viya to classify churn accurately, so the bank can identify customers most at risk of leaving.
2. Select a champion model from the tested pipelines that achieves the best performance for churn prediction.
3. Interpret the champion model to understand the main churn drivers.

1.4 Proposed Methodology

This project follows the CRISP-DM methodology. It starts with business understanding, focused on reducing churn and improving retention. The data understanding phase involves exploring the data set and checking for missing values or outliers. Data preparation includes cleaning, imputing missing values, and transforming variables for modeling. Modeling is carried out in SAS Viya using algorithms such as Gradient Boosting and Random Forest, with multiple hypertuned pipelines tested. Evaluation is performed using metrics such as accuracy, misclassification rate, Area Under the ROC Curve (AUC), Gini coefficient, and lift to select a champion model. The final step is the interpretation of the champion model, using both global and local explanations to identify important churn drivers. Although deployment is outside of the scope of this assignment, the results are discussed in terms of how they could be integrated into a system to support real-time churn management.

2 Solution Development

2.1 Data Cleaning

After examining the Bank Customer Churn dataset it was found to already be consistent and free from any data quality issues. Across the 10,000 observations and 13 features, no missing values and no duplicate records were present. The variables also displayed no anomalies that required cleaning. As a result, no imputation, outlier treatment, or record adjustment was required. Confirming the dataset was already ready for analysis and could be used directly in exploratory data analysis and model preparation without any additional cleaning.

2.2 Project Settings

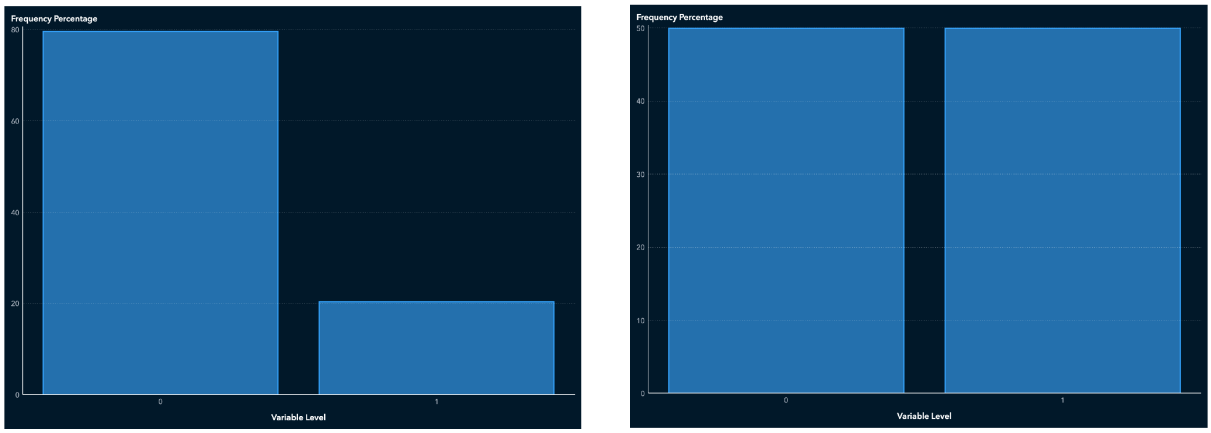
Table 2: Partitioning, Sampling, and Rules Settings in SAS Viya

Category	**Configuration**
Partitioning	Stratified split: 60% training, 20% validation, 20% test
Sampling	Event-based sampling enabled; 50% churn, 50% non-churn
Rules	KS statistic for class selection; ASE for interval selection

The dataset was partitioned using a stratified method, assigning 60% of records to training, 20% to validation, and 20% to testing. Stratification ensures the proportional representation of churn and non-churn cases across the splits. Event-based sampling was enabled to balance the dataset so that churn and non-churn classes were normally distributed. This was necessary as the original dataset only contained 20.37% churn cases, which had the potential to bias models toward predicting the majority non-churn class.

The Kolmogorov–Smirnov (KS) statistic was selected as the class selection measure, as it evaluates the degree of separation between churn and non-churn distributions. This is preferred in churn prediction tasks, where the goal is not only to classify customers correctly but also to rank them according to their probability of leaving the bank. KS therefore provides a more meaningful measure of model discrimination compared to simple accuracy.

2.3 EDA Findings and Interpretation



(a) Original churn distribution (20%). (b) Balanced churn distribution (50/50).

Figure 1: Comparison of churn distributions before and after event-based sampling.

The comparison in Figure 1 shows how event-based sampling affected the dataset. Originally, churn accounted for only about 20% of customers, which could bias models toward predicting non-churn. Balancing the classes ensured a 50/50 split, providing the models a fairer representation of both churn and non-churn.

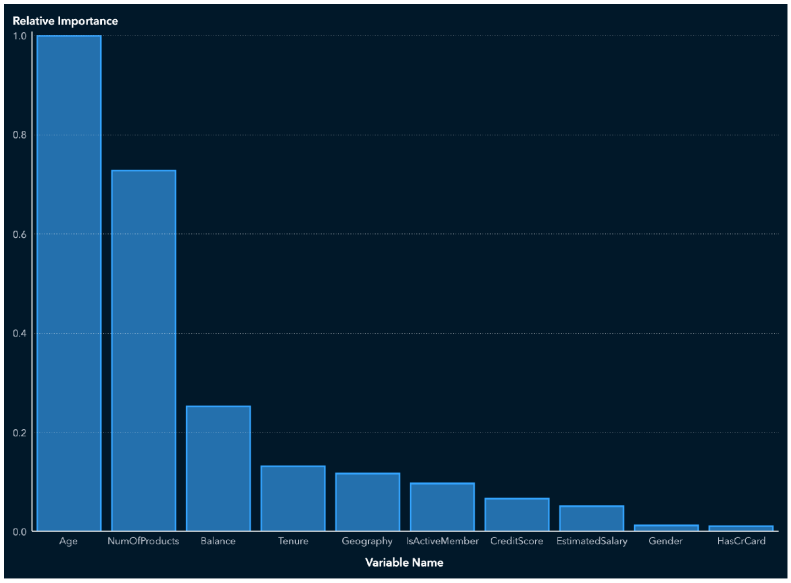
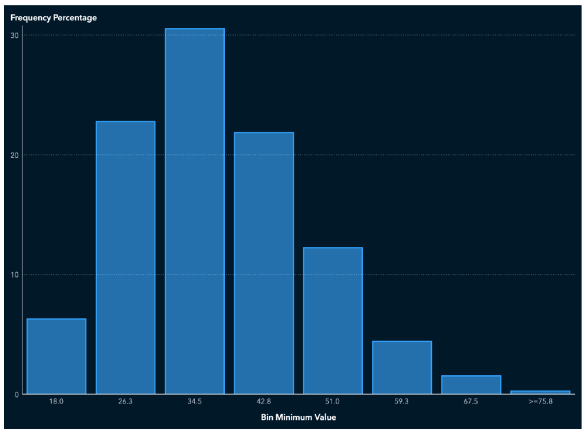
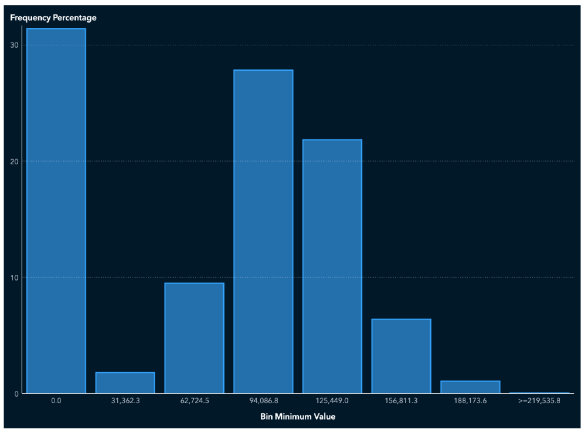


Figure 2: Variable importance ranking from SAS Viya Data Exploration node.

As shown in Figure 2, the most influential variables for churn prediction were Age, Number of Products, and Balance. These variables illustrates the importance of demographic and financial characteristics, indicating that both a customer's profile and product usage strongly influence retention.



(a) Distribution of Age.



(b) Distribution of Balance.

Figure 3: Distributions of Age and Balance in the balanced dataset.

Figure 3 shows that most customers are between the ages of 30 and 45, a demographic considered more mobile in their banking relationships. Balance displayed a wide spread, with many customers having empty balances with others having very high holdings. These patterns underscore the importance of both variables.

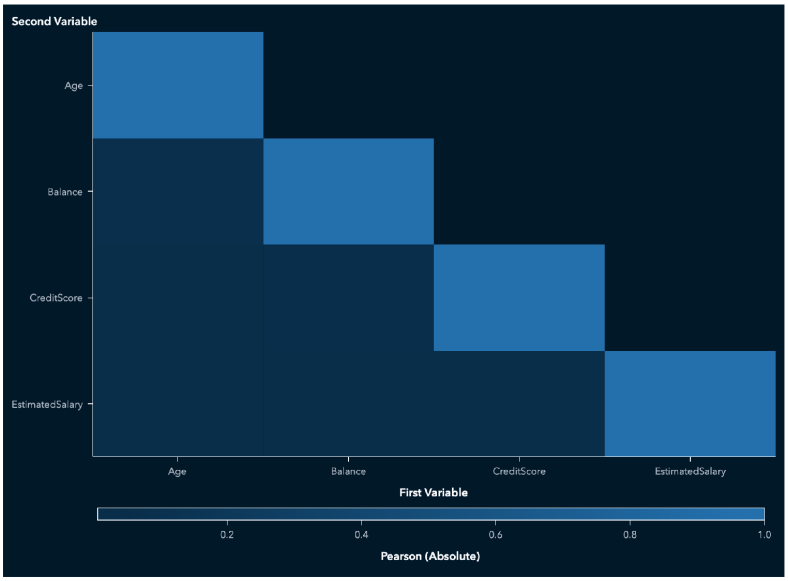


Figure 4: Pearson correlations between interval variables (absolute values).

Figure 4 confirms that the interval variables are only weakly correlated with each other (absolute Pearson values < 0.05). This reveals minimal multicollinearity, which means that each variable contributed independently to the analysis.

Table 3: Summary statistics of interval variables in the balanced dataset (n = 4,074).

Variable	Min	Max	Mean	SD	Skew	Kurtosis	Rel. Var.
Age	18	84	41.13	10.74	0.61	0.28	0.26
Balance	0	250,898	83,172.02	61,328.80	-0.31	-1.32	0.74
CreditScore	350	850	646.79	98.26	-0.08	-0.37	0.15
EstimatedSalary	11.58	199,909	101,291.25	57,389.99	-0.03	-1.19	0.57

Note. Results generated from SAS Viya Data Exploration node on the balanced dataset. Relative variability values highlight that Balance and Salary have greater spread compared to Age and CreditScore.

Table 3 shows that the interval variables behave well, with realistic ranges and only mild skewness. Age averages around 41 years, while Balance and EstimatedSalary display high variability, which is consistent with the financial diversity of the customer base. CreditScore is closely distributed around its mean, possibly meaning a more firm profile. The descriptive statistics support the adequacy of the variables for predictive modeling, with little concern about extreme skew or instability.

2.4 Data Preparation

2.4.1 Imputation

Table 4: Missing value checks for input variables (balanced dataset, n = 4,074).

Input Variable	Variable Level	Missing Values	Percent Missing	Imputable
Age	Interval	0	0.0	No
Balance	Interval	0	0.0	No
CreditScore	Interval	0	0.0	No
EstimatedSalary	Interval	0	0.0	No
Gender	Binary	0	0.0	No
Geography	Nominal	0	0.0	No
HasCrCard	Binary	0	0.0	No
IsActiveMember	Binary	0	0.0	No
NumOfProducts	Nominal	0	0.0	No
Tenure	Nominal	0	0.0	No

Note. All variables contained 0% missing values, confirming that no imputation was required. This validated the dataset for direct use in further preparation and modeling.

Table 4 shows that no variable required imputation, as there were no missing values. This reduces the risk of bias from replacements and simplified the preparation process.

2.4.2 Interactive Grouping

Table 5: Interactive Grouping results (Validation dataset, n = 4,074).

Variable	Grp	Bin Label	Events	Non-Events	WOE	Event Rate	Upper Bound
Age	1	_MISSING_	0	0	0.0	0.0	–
Age	2	Age ¡ 33	47	129	1.0097	0.1155	33
Age	3	33 ¡= Age ¡ 40	64	153	0.8716	0.1572	40
Age	4	40 ¡= Age ¡ 48	132	81	-0.4884	0.3243	48
Age	5	48 ¡= Age	164	44	-1.3157	0.4029	–
Balance	1	_MISSING_	0	0	0.0	0.0	–
Balance	2	Balance ¡ 0	0	0	0.0	0.0	0
Balance	3	0 ¡= Balance ¡ 104469.58	182	224	0.2076	0.4472	104,469.5800
Balance	4	104469.58 ¡= Balance ¡ 128876.71	103	96	-0.0704	0.2531	128,876.7100
Balance	5	128876.71 ¡= Balance	122	87	-0.3381	0.2998	–
CreditScore	1	_MISSING_	0	0	0.0	0.0	–
CreditScore	2	CreditScore ¡ 579	106	87	-0.1975	0.2604	579
CreditScore	3	579 ¡= CreditScore ¡ 645	95	102	0.0711	0.2334	645
CreditScore	4	645 ¡= CreditScore ¡ 714	109	117	0.0708	0.2678	714
CreditScore	5	714 ¡= CreditScore	97	101	0.0404	0.2383	–
EstimatedSalary	1	_MISSING_	0	0	0.0	0.0	–
EstimatedSalary	2	EstimatedSalary ¡ 50000	108	105	-0.0231	0.5070	50000
EstimatedSalary	3	50000 ¡= Salary ¡ 100000	94	102	0.0423	0.4796	100000
EstimatedSalary	4	100000 ¡= Salary ¡ 150000	97	103	0.0163	0.4850	150000
EstimatedSalary	5	150000 ¡= Salary	97	97	0.0	0.5	–
Geography	1	France	184	200	0.0790	0.4792	–
Geography	2	Germany	152	113	-0.3016	0.5737	–
Geography	3	Spain	60	94	0.4140	0.3896	–
Gender	1	Female	168	127	-0.2602	0.5695	–
Gender	2	Male	228	280	0.1903	0.4489	–
HasCrCard	1	No	183	170	-0.1069	0.5184	–
HasCrCard	2	Yes	213	237	0.0874	0.4730	–
IsActiveMember	1	No	292	211	-0.3135	0.5803	–
IsActiveMember	2	Yes	104	196	0.5368	0.3465	–
NumOfProducts	1	1	156	305	0.6682	0.3382	–
NumOfProducts	2	2	176	102	-0.4829	0.6338	–
NumOfProducts	3	3	47	0	-15.9417	1.0	–
NumOfProducts	4	4	17	0	-15.9417	1.0	–
Tenure	1	0	51	56	0.0776	0.4766	–
Tenure	2	1	53	52	-0.0309	0.5048	–
Tenure	3	2	58	57	-0.0096	0.5043	–
Tenure	4	3	58	58	0.0	0.5	–
Tenure	5	4	52	53	0.0094	0.4952	–
Tenure	6	5	61	56	-0.0428	0.5217	–
Tenure	7	6	46	57	0.2079	0.4466	–
Tenure	8	7	40	51	0.2022	0.4396	–
Tenure	9	8	29	48	0.4901	0.3766	–
Tenure	10	9	21	47	0.7836	0.3090	–
Tenure	11	10	27	44	0.5112	0.3803	–
Tenure	12	11	18	47	0.9265	0.2778	–
Tenure	13	12	13	47	1.1996	0.2167	–

Note. The Interactive Grouping node in SAS Viya binned both interval variables (e.g., Age, Balance, CreditScore, EstimatedSalary) and categorical variables (e.g., Gender, Geography, HasCrCard, IsActiveMember, NumOfProducts, Tenure). Each bin is associated with its Weight of Evidence (WOE), event rate, and cut-off.

Table 5 shows that Age has a clear increase in churn risk across bins, while Balance, CreditScore, and EstimatedSalary have moderate differences. Nominal predictors such as Geography and Gender were also grouped into GRP_ and WOE_ forms to improve interpretability.

2.4.3 Transformations

Table 6: Transformation checks for input variables (balanced dataset, n = 4,074).

Variable	**Variable Level**	**Transformation Applied**
Age	Interval	None required
Balance	Interval	None required
CreditScore	Interval	None required
EstimatedSalary	Interval	None required
Gender	Binary	None required
Geography	Nominal	None required
GRP_Age	Ordinal	Derived via Interactive Grouping
GRP_Geography	Ordinal	Derived via Interactive Grouping
GRP_IsActiveMember	Ordinal	Derived via Interactive Grouping
GRP_NumOfProducts	Ordinal	Derived via Interactive Grouping
HasCrCard	Binary	None required
IsActiveMember	Binary	None required
NumOfProducts	Nominal	None required
Tenure	Nominal	None required
WOE_Age	Interval	Derived via Weight of Evidence
WOE_Geography	Interval	Derived via Weight of Evidence
WOE_IsActiveMember	Interval	Derived via Weight of Evidence
WOE_NumOfProducts	Interval	Derived via Weight of Evidence

Note. SAS Viya's Transformations node evaluated the distributional properties of each input variable (e.g., skewness, kurtosis, variability). Since all variables distributions were acceptable, no transformations such as square, square root, or logarithmic adjustments were used. In addition, the Interactive Grouping and Weight of Evidence (WOE) automatically generated grouped (GRP_) and evidence-weighted (WOE_) variables for selected predictors.

Table 6 shows that no original variable required any kind of transformation, while the grouped and WOE variables were created by SAS Viya to capture any non-linear patterns and categorical splits. This guarantees that the final dataset includes both clean predictors and customised versions to hopefully boost model performance.

2.4.4 Feature Machine

Table 7: Engineered features generated by SAS Viya Feature Machine.

Input Variable	New Feature	Transformation Method	Ranking Criterion	Rank
Age	nhoks_nloks_dtree_5_Age	five bin decision tree binning	0.0941	1
Age	nhoks_nloks_pow_n1_Age	power(-1) + impute(median)	0.0752	2
Balance	cpy_int_med_imp_Balance	median imputation	0.0126	1
CreditScore	cpy_int_med_imp_CreditScore	median imputation	0.0058	1
CreditScore	all_l_oks_dtree_10_CreditScore	ten bin decision tree binning	0.0016	2
EstimatedSalary	cpy_int_med_imp_EstimatedSalary	median imputation	0.0042	1
EstimatedSalary	all_l_oks_dtree_10_var_1_	ten bin decision tree binning	0.0023	2
GRP_Age	cpy_nom_mode_imp_lab_GRP_Age	mode imputation + label transformation	0.0773	1
GRP_Geography	cpy_nom_mode_imp_lab_var_2_	mode imputation + label transformation	0.0228	1
GRP_IsActiveMember	cpy_nom_mode_imp_lab_var_3_	mode imputation + label transformation	0.0257	1
GRP_NumOfProducts	cpy_nom_mode_imp_lab_var_4_	mode imputation + label transformation	0.1182	1
Gender	cpy_nom_mode_imp_lab_Gender	mode imputation + label transformation	0.0147	1
Geography	cpy_nom_mode_imp_lab_Geography	mode imputation + label transformation	0.0228	1
IsActiveMember	cpy_nom_mode_imp_lab_var_5_	mode imputation + label transformation	0.0257	1
NumOfProducts	cpy_nom_mode_imp_lab_var_6_	mode imputation + label transformation	0.1182	1
Tenure	cpy_nom_mode_imp_lab_Tenure	mode imputation + label transformation	0.0472	1
WOE_Age	cpy_int_med_imp_WOE_Age	median imputation	0.0089	1
WOE_Geography	cpy_int_med_imp_var_9_	median imputation	0.0076	1
WOE_IsActiveMember	cpy_int_med_imp_var_10_	median imputation	0.0052	1
WOE_NumOfProducts	cpy_int_med_imp_var_11_	median imputation	0.0063	1
Age	cpy_int_med_imp_Age	median imputation	0.0213	1
Balance	nhoks_nloks_dtree_5_Balance	five bin decision tree binning	0.0037	2
Balance	nhoks_nloks_pow_n1_Balance	power(-1) + impute(median)	0.0031	3
CreditScore	nhoks_nloks_dtree_5_CreditScore	five bin decision tree binning	0.0047	3
CreditScore	nhoks_nloks_pow_n1_CreditScore	power(-1) + impute(median)	0.0029	4
EstimatedSalary	nhoks_nloks_dtree_5_var_1_	five bin decision tree binning	0.0024	3
EstimatedSalary	nhoks_nloks_pow_n1_var_1_	power(-1) + impute(median)	0.0019	4
GRP_Age	all_l_oks_dtree_10_GRP_Age	ten bin decision tree binning	0.0048	2
GRP_Geography	all_l_oks_dtree_10_var_2_	ten bin decision tree binning	0.0027	2
GRP_IsActiveMember	all_l_oks_dtree_10_var_3_	ten bin decision tree binning	0.0036	2
GRP_NumOfProducts	all_l_oks_dtree_10_var_4_	ten bin decision tree binning	0.0064	2
Tenure	nhoks_nloks_dtree_5_Tenure	five bin decision tree binning	0.0059	2
Tenure	nhoks_nloks_pow_n1_Tenure	power(-1) + impute(median)	0.0043	3

Note. The Feature Machine node in SAS Viya automatically engineered 34 features using methods such as decision tree binning, power transformations, and imputations.

Table 7 displays how Feature Machine expanded the dataset with engineered variables, particularly from Age, Balance, and CreditScore. These additional features capture non-linear and interaction effects that can improve model prediction.

2.4.5 Variable Selection

Table 8: Variables retained and rejected after Variable Selection (balanced dataset, n = 4,074).

Variable	Level	Role	Reason
EXITED	BINARY	TARGET	–
ALL_L_OKS_DTREE_10_CREDITSCORE	NOMINAL	INPUT	–
CPY_INT_MED_IMP_VAR_7_	INTERVAL	INPUT	–
CPY_INT_MED_IMP_VAR_8_	INTERVAL	INPUT	–
CPY_INT_MED_IMP_WOE_AGE	INTERVAL	INPUT	–
CPY_NOM_MODE_IMP_LAB_GENDER	NOMINAL	INPUT	–
CPY_NOM_MODE_IMP_LAB_VAR_4_	NOMINAL	INPUT	–
HS_BC_N2_WOE_GEOGRAPHY	INTERVAL	INPUT	–
NHOKS_NLOKS_DTREE_5_AGE	NOMINAL	INPUT	–
CUSTOMERID	INTERVAL	ID	–
SURNAME	NOMINAL	ID	–
DMINDEX	NOMINAL	KEY	–
PARTIND	NOMINAL	PARTITION	–
ALL_L_OKS_DTREE_10_VAR_1_	NOMINAL	REJECTED	Variance Explained (Supervised)
CPY_INT_MED_IMP_BALANCE	INTERVAL	REJECTED	Variance Explained (Supervised)
CPY_INT_MED_IMP_CREDITSCORE	INTERVAL	REJECTED	Variance Explained (Supervised)
CPY_INT_MED_IMP_ESTIMATEDSALARY	INTERVAL	REJECTED	Variance Explained (Supervised)
CPY_INT_MED_IMP_WOE_GEOGRAPHY	INTERVAL	REJECTED	Variance Explained (Supervised)
CPY_NOM_MODE_IMP_LAB_GEOGRAPHY	NOMINAL	REJECTED	Variance Explained (Supervised)
CPY_NOM_MODE_IMP_LAB_GRP_AGE	NOMINAL	REJECTED	Variance Explained (Supervised)
CPY_NOM_MODE_IMP_LAB_TENURE	NOMINAL	REJECTED	Variance Explained (Supervised)
CPY_NOM_MODE_IMP_LAB_VAR_2_	NOMINAL	REJECTED	Variance Explained (Supervised)
CPY_NOM_MODE_IMP_LAB_VAR_3_	NOMINAL	REJECTED	Variance Explained (Supervised)
CPY_NOM_MODE_IMP_LAB_VAR_5_	NOMINAL	REJECTED	Variance Explained (Supervised)
CPY_NOM_MODE_IMP_LAB_VAR_6_	NOMINAL	REJECTED	Variance Explained (Supervised)
HASCRCARD	BINARY	REJECTED	Variance Explained (Supervised)
HS_BC_0_WOE_GEOGRAPHY	INTERVAL	REJECTED	Variance Explained (Supervised)
HS_BC_0_WOE_ISACTIVEMEMBER	INTERVAL	REJECTED	Variance Explained (Supervised)
HS_BC_N1_WOE_GEOGRAPHY	INTERVAL	REJECTED	Variance Explained (Supervised)
HS_BC_N1_WOE_ISACTIVEMEMBER	INTERVAL	REJECTED	Variance Explained (Supervised)
HS_BC_N2_WOE_ISACTIVEMEMBER	INTERVAL	REJECTED	Variance Explained (Supervised)
HS_BC_P1_WOE_GEOGRAPHY	INTERVAL	REJECTED	Variance Explained (Supervised)
HS_BC_P1_WOE_ISACTIVEMEMBER	INTERVAL	REJECTED	Variance Explained (Supervised)
HS_BC_P2_WOE_GEOGRAPHY	INTERVAL	REJECTED	Variance Explained (Supervised)
HS_BC_P2_WOE_ISACTIVEMEMBER	INTERVAL	REJECTED	Variance Explained (Supervised)
HS_DTREE_DISCT10_VAR_7_	NOMINAL	REJECTED	Variance Explained (Supervised)
HS_DTREE_DISCT10_WOE_GEOGRAPHY	NOMINAL	REJECTED	Variance Explained (Supervised)
HS_DTREE_DISCT5_VAR_7_	NOMINAL	REJECTED	Variance Explained (Supervised)
HS_DTREE_DISCT5_WOE_GEOGRAPHY	NOMINAL	REJECTED	Variance Explained (Supervised)
NHOKS_NLOKS_POW_N1_AGE	INTERVAL	REJECTED	Variance Explained (Supervised)

Note. The Variable Selection node retained essential target and input features while excluding ID fields and rejecting variables with low predictive contribution.

Table 8 shows all retained and rejected variables. This transparent selection process ensured that the final dataset emphasized predictors with the highest relevance for modeling churn.

2.5 Modeling

2.5.1 Baseline Model: Logistic Regression

Logistic regression is used as the baseline model because it is the default classification template in SAS Viya and provides a clear reference point for evaluating more advanced models. It is widely applied for churn prediction due to its interpretability, ability to estimate the probability of churn, and proficiency on structured tabular data. However, its performance is limited when complex non-linear relationships are present between predictors and churn, which justifies the need to compare it against ensemble models such as Random Forest and Gradient Boosting.

Table 9: Baseline performance of Logistic Regression model (balanced dataset, $n = 816$).

Metric	**Value**
KS (Youden)	0.5417
Accuracy	0.7010
Average Squared Error (ASE)	0.2098
Area Under ROC (AUC)	0.8434
Cumulative Lift	1.9853
Cumulative Captured Response (%)	19.85
Cutoff	0.50
F1 Score	0.5523
False Positive Rate (FPR)	0.0898
Misclassification Rate	0.2990
Gini Coefficient	0.6868

Note. Results are taken from the Model Comparison node in SAS Viya. Metrics highlight overall classification accuracy, discriminatory power (AUC, KS, Gini), and error rates. Logistic regression serves as the baseline model for comparison against more advanced approaches.

The baseline logistic regression achieved an AUC of 0.84 and a Gini of 0.69, a good discriminatory ability between churned and retained customers. Its accuracy of 70% shows a reasonable balance, but the F1 score of 0.55 suggests that identifying churners is more difficult than predicting non-churn. Although the model establishes a good benchmark, the relatively high misclassification rate (29.9%) shows the need for a stronger predictive model.

2.5.2 Gradient Boosting

Gradient Boosting was selected as it is an effective machine learning model that predicts customer churn. It achieves this by building a series of decision trees that learn from the mistakes of the previous trees. This sequential process allows it to be more accurate than other models like logistic regression and Random Forest, particularly for complex problems with non-linear relationships. Its flexibility and resistance to overfitting further make it a solid choice for identifying customers at risk of leaving.

Table 10: Gradient Boosting hyperparameter settings (baseline vs tuned runs).

Parameter	**Baseline GB**	**GB (1)**	**GB (2)**	**GB (3)**
Number of trees	100	200	500	1000
Learning rate	0.10	0.10	0.05	0.02
Subsample rate	0.50	0.70	0.80	0.90
L1 regularization	0.00	0.00	0.01	0.00
L2 regularization	1.00	0.10	0.05	0.01
Maximum depth	4	4	6	8
Minimum leaf size	5	20	10	5
Number of interval bins	50	30	30	50
Stagnation	5	20	30	50
Tolerance	0	0.001	0.0005	0.0002

Note. Default Gradient Boosting parameters were incrementally tuned across three runs. GB (1) achieved the best trade-off between depth, regularization, and learning rate, leading to its selection as the Champion model.

Table 11: Performance comparison of Gradient Boosting models (balanced dataset, $n = 816$).

Metric	**Baseline GB**	**GB (1) Champion ★**	**GB (2)**	**GB (3)**
KS (Youden)	0.5564	0.5637	0.5049	0.4828
Accuracy	0.6483	0.7365	0.7341	0.7194
Average Squared Error (ASE)	0.2325	0.1915	0.1987	0.2211
Area Under ROC (AUC)	0.8492	0.8532	0.8194	0.7967
Cumulative Lift	1.9853	1.9608	1.9118	1.9118
Cumulative Captured Response (%)	19.85	19.61	19.12	19.12
Cutoff	0.50	0.50	0.50	0.50
Misclassification Rate	0.3517	0.2635	0.2659	0.2806
Gini Coefficient	0.6984	0.7063	0.6387	0.5934

Note. Gradient Boosting was tested under four settings, with GB (1) emerging as the Champion model (marked with ★). It delivered the best balance between accuracy, AUC, and error reduction, making it the selected Gradient Boosting configuration for further interpretation.

Table 11 reveals that the tuned GB (1) Champion model achieved higher accuracy (73.65%) and a lower misclassification rate (26.35%) compared to the baseline at 64.83% and 35.17%. Although the cumulative lift and captured response were similiar, GB (1) showed a stronger discriminatory power with an AUC of 0.85 and Gini of 0.71, slightly surpassing the baseline. These improvements verify that tuning helped the model capture more of the subtle patterns in the data, reducing classification errors and improving separation between churners and non-churners.

2.5.3 Random Forest

Random Forest was chosen as the second ensemble model because it is widely used for classification problems and provides a strong performance by averaging multiple decision trees. Compared to Gradient Boosting, which builds trees sequentially, Random Forest builds them in parallel, which reduces the risk of overfitting and improving generalisation. It also handles large numbers of variables and interactions efficiently, making it a suitable model for churn prediction.

Table 12: Random Forest hyperparameter settings (baseline vs tuned runs).

Parameter	**Baseline RF**	**RF (1)**	**RF (2)**	**RF (3)**
Number of trees	100	500	700	500
Maximum depth	20	16	14	16
Minimum leaf size	5	10	12	12
In-bag sample proportion	0.60	0.55	0.60	0.65

Note. The Random Forest pipeline was tuned by adjusting tree count, depth, leaf size, and in-bag proportion. RF (1) achieved the best balance of accuracy and error reduction, making it the Champion model.

Table 13: Performance comparison of Random Forest models (balanced dataset, $n = 816$).

Metric	**Baseline RF**	**RF (1) Champion ★**	**RF (2)**	**RF (3)**
KS (Youden)	0.5686	0.5686	0.5662	0.5637
Accuracy	0.6949	0.6850	0.6863	0.6900
Average Squared Error (ASE)	0.2056	0.2110	0.2112	0.2104
Area Under ROC (AUC)	0.8517	0.8507	0.8502	0.8503
Cumulative Lift	1.9608	1.9853	1.9853	1.9853
Cumulative Captured Response (%)	19.61	19.85	19.85	19.85
Cutoff	0.50	0.50	0.50	0.50
Misclassification Rate	0.3051	0.3150	0.3137	0.3100
Gini Coefficient	0.7033	0.7014	0.7004	0.7007

Note. Random Forest was tested under four settings, with RF (1) selected as the Champion model (marked with ★). While its accuracy was slightly lower than the baseline, it maintained comparable discriminatory power (AUC, Gini) and demonstrated stable performance across different configurations.

Table 13 reveals that the tuned RF (1) Champion model recorded an accuracy of 68.50% and a misclassification rate of 31.50% slightly below the baseline RF at 69.49% and 30.51%. Although the baseline was slightly stronger for those two metrics, RF (1) achieved a higher cumulative lift (1.9853 vs. 1.9608) and captured response percentage (19. 85% vs 19. 61%), while keeping a comparable discriminatory power (AUC = 0.8507, Gini = 0.7014). This illustrates how RF (1) has balanced predictive performance with stronger response capture, which justified its selection as the Champion in the Random Forest pipeline and the tunining done.

3 Discussion of Model Outcomes

3.1 Pipeline Comparison

Table 14: Pipeline comparison results on the test dataset (n = 816).

Champion	Name	Algorithm Name	Pipeline Name	KS (Youden)	Number of Observations
★	Forest (1)	Forest	RandomForest	0.5686	816
	Gradient Boosting (1)	Gradient Boosting	GradientBoost	0.5637	816
	Logistic Regression	Logistic Regression	Basic	0.5417	816

Forest (1) achieved the highest KS (Youden) value (0.5686), slightly outperforming Gradient Boosting (1) and Logistic Regression. Where 68.5% of the Test partition was correctly classified using the Forest (1) model.

3.2 Model Interpretability

3.2.1 Global Interpretability: Surrogate Model Variable Importance

Table 15: Top five variables ranked by surrogate model importance.

Variable Label	Variable Name	Relative Importance	Role	Level
Age: Not high (outlier, kurtosis, skewness) – five bin decision tree binning	nhoks_nloks_dtree_5_Age	1.0000	Input	Nominal
WOE_NumOfProducts: Low missing rate – median imputation	cpy_int_med_imp_var_8_	0.9488	Input	Interval
GRP_NumOfProducts: Low missing rate – mode imputation + label transformation	cpy_nom_mode_imp_lab_var_4_	0.9488	Input	Nominal
WOE_Age: Low missing rate – median imputation	cpy_int_med_imp_WOE_Age	0.6645	Input	Interval
WOE_Geography: High skewness – Box-Cox(λ=-2) + impute(median)	hs_bc_n2_WOE_Geography	0.1107	Input	Interval

Note. Relative importance scores are scaled between 0 and 1, where higher values indicate greater contribution to churn prediction.

Based on Table 15, the most significant factors in the model's decision-making process were age-related and product-related features. These were the strongest predictors of churn. While geographic variables were also included in the top 5, their influence wasn't as strong. This implies that a customer's demographic andtheir engagement with a product are the primary drivers of churn.

3.2.2 Global Interpretability: Partial Dependence Plots

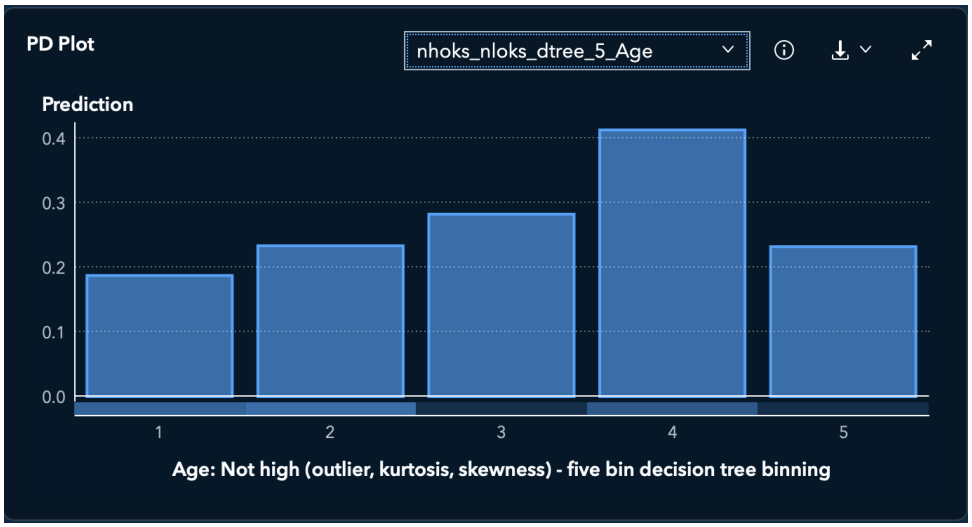


Figure 5: Partial dependence plot for Age (five-bin decision tree binning).

The highest average predicted churn probability was 0.41 when the grouped age value was 4, while the lowest was 0.19 at 1. This indicates that older age groups are more strongly associated with churn compared to younger ones.

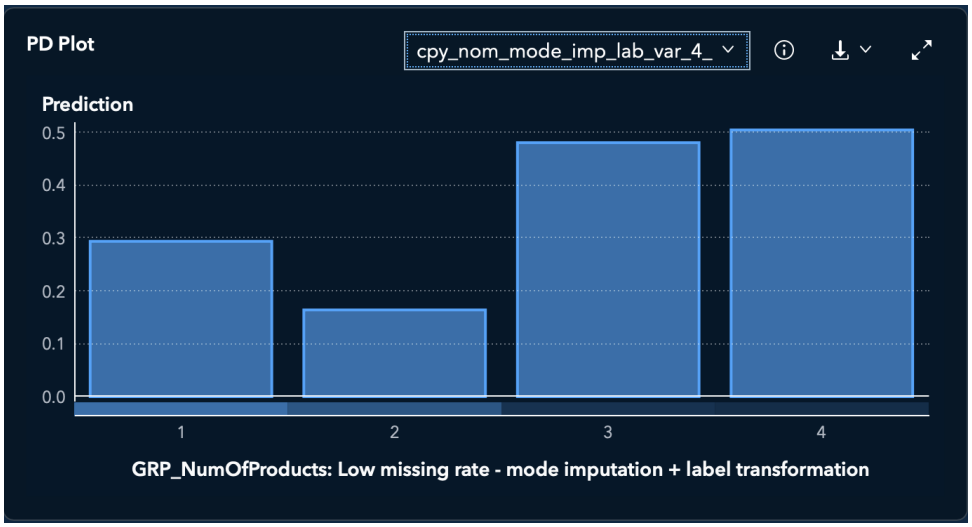


Figure 6: Partial dependence plot for GRP_NumOfProducts (mode imputation + label transformation).

The grouped number of products variable showed its highest predicted churn probability of 0.50 when equal to 4, and its lowest at 0.16 when equal to 2. Suggesting that customers with more bank products are more likely to churn, whereas those with moderate product usage are less likely to leave.

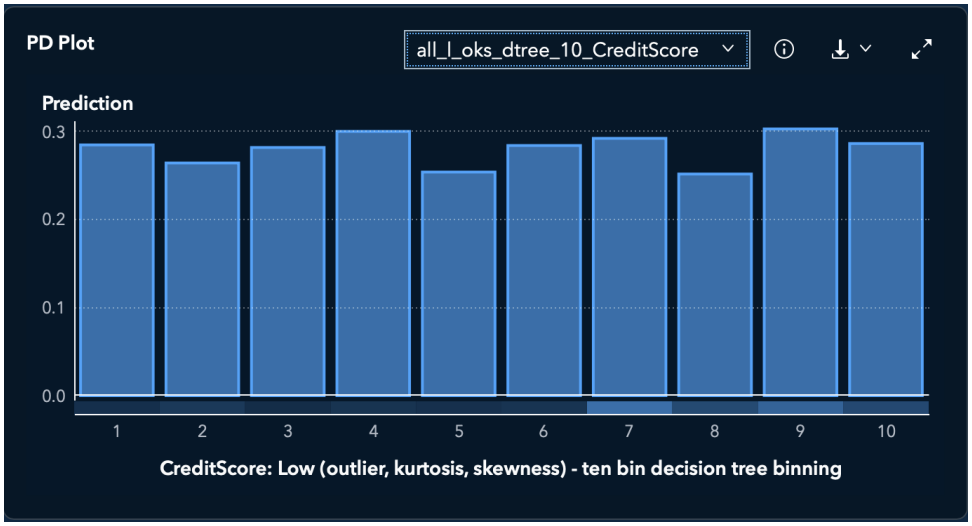


Figure 7: Partial dependence plot for CreditScore (ten-bin decision tree binning).

For the grouped credit score variable, the predicted churn probability peaked at 0.30 when the score equaled 9, while the lowest was 0.25 at 8. DIsplaying only a slight difference in churn risk across score groups, possibly meaning that credit score plays a weaker role compared to other factors.

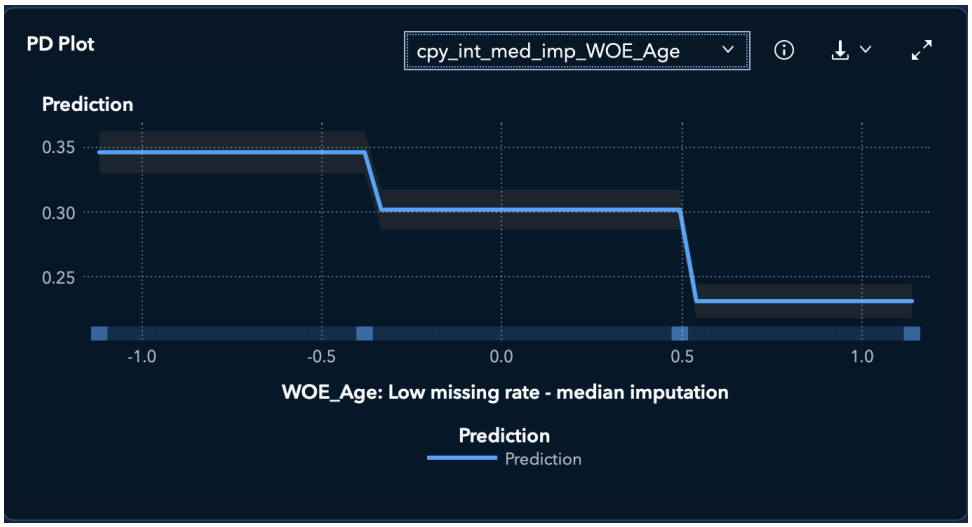


Figure 8: Partial dependence plot for WOE_Age (median imputation).

The weight-of-evidence age feature showed a maximum predicted churn probability of 0.35 at -1.121 and a minimum of 0.23 at 0.541. Suggesting that certain transformed age segments are more predictive of churn risk, revealing the nonlinear effect of age.

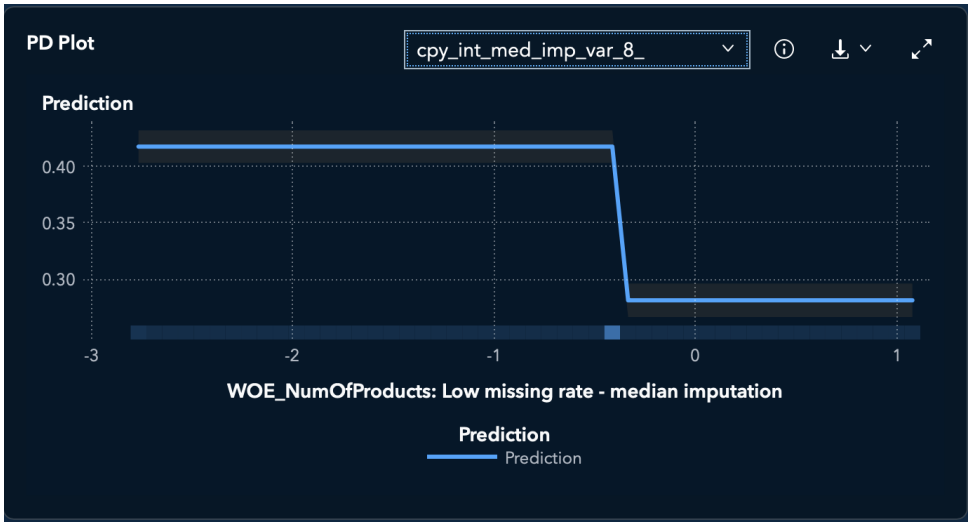


Figure 9: Partial dependence plot for WOE_NumOfProducts (median imputation).

The weight-of-evidence number of products feature showed its highest predicted churn probability of 0.42 at -2.764 and lowest at 0.28 at -0.335. This reinforces that certain product ownership patterns contribute more heavily to churn than others.

3.2.3 Local Interpretability: HyperSHAP Values

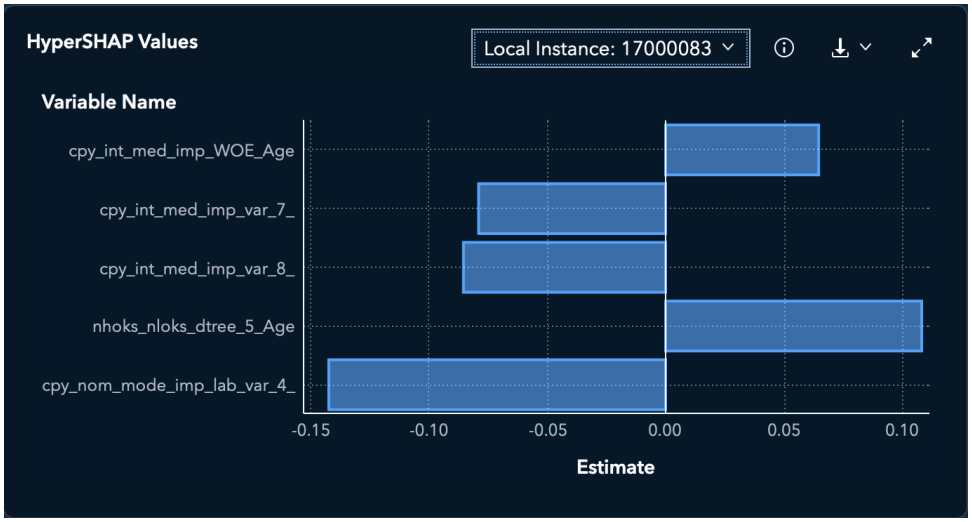


Figure 10: HyperSHAP values for Local Instance: 17000083

For Customer 17000083, the biggest factor increasing their churn risk up is their Age (binned into 5 groups), Age (WOE, median-imputed) also contributes reinforcing the effect of age. In contrast, their Number of Products (grouped, mode-imputed) lowers the risk of churn. Other factors, like a separate measure of Number of Products (WOE, median-imputed) and whether they are an IsActiveMember (WOE, median-imputed) have moderate effects that also reduce the chance of churn.

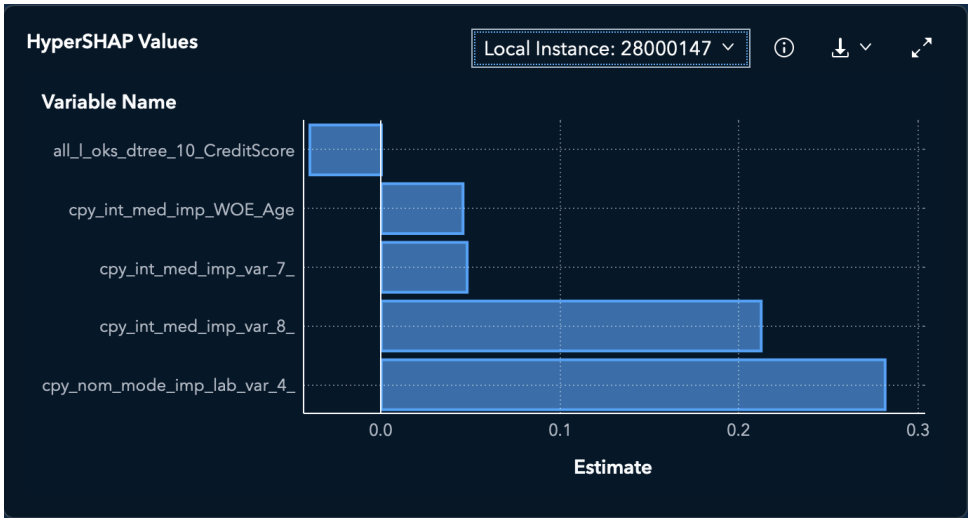


Figure 11: HyperSHAP values for Local Instance: 28000147

For Customer 28000147, their churn prediction is mainly motivated by their Number of Products (grouped, mode-imputed) and a separate measure of Number of Products (WOE, median-imputed), which both significantly increase their risk of churn. Factors like Age (WOE, median-imputed) and IsActiveMember (WOE, median-imputed) also contribute, but not as significantly. In contrast, Credit Score (binned into 10 groups) lowers the risk of churn.

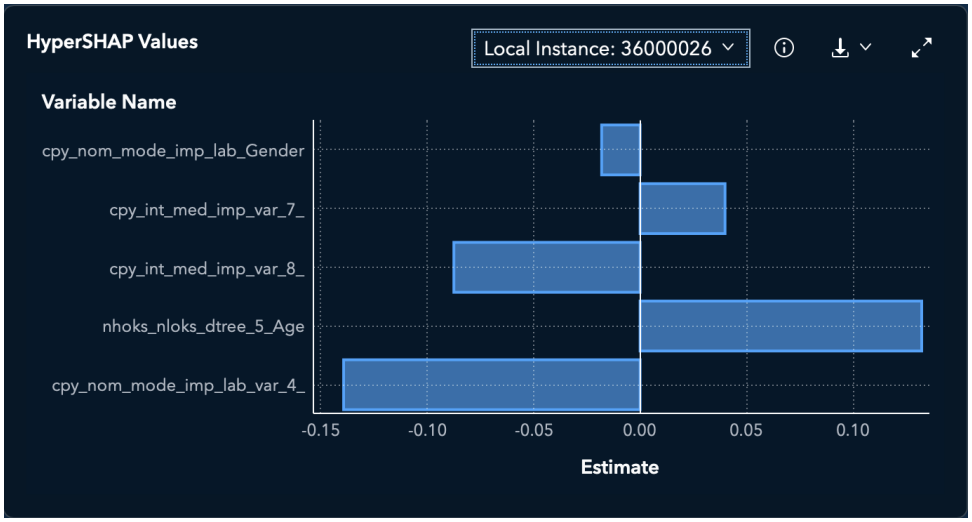


Figure 12: HyperSHAP values for Local Instance: 36000026

For Customer 36000026, the prediction is a balance between two factors. Their Age (binned into 5 groups) is a strong factor that increases their churn risk, while their Number of Products (grouped, mode-imputed) works to reduce it. Other factors that reduce the risk is from their Number of Products (WOE, median-imputed) and Their Gender (mode-imputed with label transform). However, IsActiveMember (WOE, median-imputed) adds pressure, raising churn risk.

3.3 Business Alignment

This solution addresses the bank's problem of not knowing which customers are susceptible to churn. By using Random Forest (The champion model), we have created a reliable way to predict the banks churn. Meaning the bank can now proactively keep customers instead of only reacting when they have already left.

The model also revealed that age, the number of products a customer has, and their activity status are significant factors to predict churn. This provides the bank practical and data-driven insights. Rather than generic campaigns, the bank has a clear approach for their retention strategies which could include.

- **Age Specific Engagement:** Target older customers with tailored retention programs, like senior-specific loyalty rewards or dedicated advisory services. The insights shows that this demographic has a higher churn risk so they need a more specialised approach to feel valued and prevent churn.
- **Optimize Product Portfolio:** Instead of pushing customers to have more and more products, focus on ensuring that they get value from the ones they already have. The model shows that both too few and too many products can be risky. The key is to help customers make the most their existing accounts efficiently, which will improve thier long-term engagement and reduce churn.
- **Proactive Outreach to Inactive Members:** Inactivity is a strong warning sign for churn. Focus on targeted campaigns like personalised usage-based rewards to re-engage at risk customers.

With both a strong predictive model and these actionable insights, the bank can now forecast churn with confidence and use data to make smarter decisions that improve customer retention and long-term value.

References

de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: a machine learning approach. *Neural Computing and Applications*, *34*(14), 11751–11768. https://doi.org/10.1007/s00521-022-07067-x