

Group – 7

SAS Group Project Report

Name	Student ID
Abhik Sarkar	301333781
Nimish Abraham	301312186
Vinay Beesa Gnaneshwar	301335423

Dataset- Bike Buyers

<https://www.kaggle.com/datasets/heeraldedhia/bike-buyers>

Contents

Introduction	3
Data Setup & Exploration.....	4
Variables and changes made	5
Maximal Tree	6
Replace/Filter.....	8
Maximal Tree	9
Misclassification Tree.....	10
Average Square Error Tree.....	12
Competing Splits for ASE Tree	13
Impute.....	14
Skewness.....	15
Full Regression	15
Forward Regression	16
Backward Regression	16
Stepwise Regression	17
Interpretation of Regression (forward)	18
Interpretations of the Odds Ratio Estimates are as follows:.....	19
Full Neural Network (from impute)	21
Iteration Plot	22
Optimal Neural Network (forward)	22
Assessment (model comparison).....	23
ROC Chart.....	23
Lift Chart.....	25
Conclusion.....	26
References	27

Introduction

The global bicycle market has experienced significant growth over the past decade, primarily driven by an increasing awareness on health, nutrition and overall wellbeing by consumers. Adding to this success is the advancement in bike technology and a surge in bicycle manufacturing infrastructure and its R&D. The COVID-19 pandemic too has acted as a catalyst in the spike in bicycle demand, as people sought safe and socially distanced means of transportation, all being beneficial to their overall health and wellbeing. According to data from Statista, in 2024 the projected revenue of the bicycle market across the globe is set to reach a whopping US\$62.08 billion (Statista, p.1).

Having a thorough understanding of the intricacies behind such consumer purchasing decisions is extremely key to our strategic business planning. This research dives into the myriad of factors that influence bicycle purchasing, examines how personal preferences, financial conditions, and socio-demographic factors affect in shaping the decision of purchasing.

Our study focuses on uncovering the motivations behind bicycle purchases, taking into account income levels, age, lifestyle choices, geography and much more. Bicycles have evolved from humble modes of transportation to symbols of healthy living and environmental sustainability. They cater to diverse needs, from transportation, leisure fitness, and extreme competition. Recent studies reveal a notable trend wherein individuals who regularly engage in biking activities often lead a health conscious lifestyle and prioritize environmentally friendly choices.

In this study, we employ predictive models to explore consumer behaviors in bicycle purchases, utilizing a dataset that is filled with all necessary information to carry out said analysis. The dataset spans Europe, North America, and the Pacific region, which are the prime markets for bicycle sales globally. Our approach includes advanced modelling techniques like decision trees, regressions, and neural networks, all executed using the SAS Enterprise Miner platform.

The objective of the study is to assess the predictive power of these models in forecasting bicycle purchasing behaviors. By analyzing variables such as income levels, geographic locations, age, preferences, distance of commute, education, gender, marital status, occupation and their previous history of owning a bicycle, we aim to predict the likelihood of an individual becoming a bicycle buyer based on their profile. Additionally, we seek to identify the best markets, demographic segments and income groups to target selling of bicycles.

To achieve this, we employed robust data setup and exploration techniques to ensure that the variables were appropriate for the analysis. We ran multiple data modelling techniques to assess and compare their performance. Metrics such as the ROC index and the ASE values were used to evaluate the effectiveness of each model, enabling us in determining which model excels at predicting bicycle purchasing behaviors.

Our findings offer valuable insights which can be then translated into strategic decisions, helping in optimizing market efforts in bicycle sales across Europe, North America, and the Pacific. By identifying the most promising markets and demographic segments, we can then tailor our strategies to effectively target and engage the potential bicycle buyers, yielding in a growth in sales.

Data Setup & Exploration

The dataset used in this analysis comprises information on 1,000 users and their respective attributes related to bike purchases. The dataset columns encompass:

Variables	Model Role	Measurement Level	Description
ID	ID	Nominal	Unique identifier for everyone
Marital Status	Input	Nominal	Relationship status of the individual
Gender	Input	Nominal	Gender identity of the individual
Income	Input	Interval	Income level of the individual
Children	Cars	Input	Nominal
Education	Input	Ordinal	Educational background
Occupation	Input	Nominal	Professional field of the individual
Homeowner	Input	Nominal	Property ownership status
Cars	Input	Nominal	Number of cars owned
Commute Distance	Cars	Input	Nominal
Region	Input	Nominal	Geographic area of residence
Age	Input	Interval	Age of the individual
Purchased Bike	Target	Binary	(1 for Yes, 0 for No)

During the project planning stage, we experimented with two different methods of data input to determine the most effective approach of our analysis. The first method involved feeding the raw dataset into SAS and running the project as is. The second method aimed at reducing the number of variables by altering the “Commute” variable in the xls. We had to calculate the mid-point of the “Commute” distance and let SAS run the model with the amended dataset.

While the second method yielded a smaller maximal tree, we found that the ASE (Average Squared Error) scores to be much better in the first method. Taking into consideration the better ASE scores, we decided to proceed with the first method for our analysis. This approach ensured more accurate model performance and allowed us to achieve a reliable set of results.

Below is a snippet of the amendments that were made in the second method, for viz. sake.

Commute Distance	Mid-Point Commute
0-1 Miles	0.5
0-1 Miles	0.5
2-5 Miles	3.5
5-10 Miles	7.5
0-1 Miles	0.5
1-2 Miles	1.5
0-1 Miles	0.5
0-1 Miles	0.5
5-10 Miles	0.5
0-1 Miles	7.5
1-2 Miles	0.5
10+ Miles	1.5
0-1 Miles	5

Variables and changes made

Table of Variable Levels:

Level Type	Description
Interval	Numeric values with equal intervals, representing quantities with meaningful distances but no true zero point (e.g., temperature, time).
Nominal	Categorical data without inherent order. Each value signifies a distinct category (e.g., types of cars, colours, gender).
Ordinal	Categorical data with ranked order. Order matters but intervals between values may vary (e.g., education levels, satisfaction ratings, socioeconomic status).

Table of Changes:

Variable	Original Level	New Level	Description
Cars	Interval	Nominal	Changed from a continuous numerical value to distinct categories.
Education	Nominal	Ordinal	Transformed to reflect an ordered hierarchy among categories.
Commute	Interval	Nominal	Converted from a continuous value to distinct categories.

Variable Changes and Implications:

- Cars:
 - Change: Altered from interval level to nominal level.
 - Implication: Number of cars now treated as categorical, not ordered by quantity.
- Education:
 - Change: Transformed from nominal to ordinal.
 - Implication: Education levels now ordered by hierarchy, reflecting advancement.
- Commute Distance:
 - Change: Converted from interval level to nominal level.
 - Implication: Commute distances treated categorically, enhancing analysis clarity.

These changes align data with analysis needs, ensuring accuracy and insight depth.

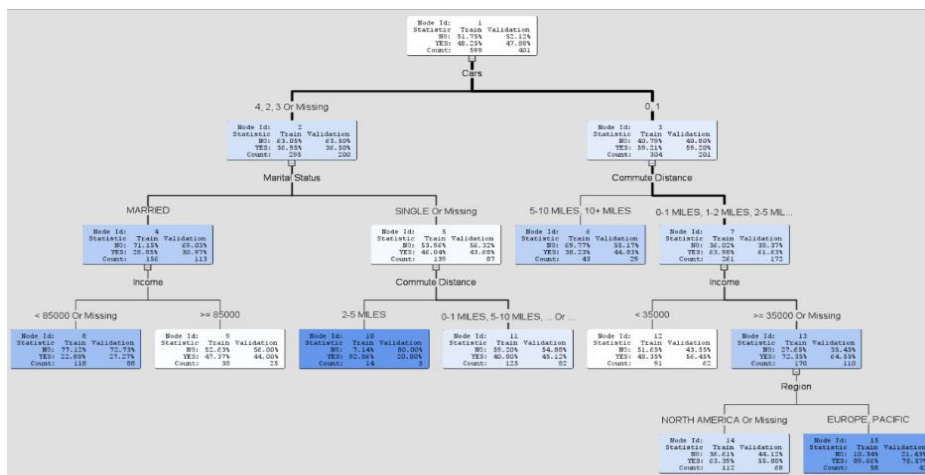
Maximal Tree

A maximal tree refers to a decision tree that's grown without any constraints or pruning. The maximal tree, while capturing intricate patterns within the training data, tends to be overly complex and can easily over fit. Overfitting occurs when a model learns the specific details and noise in the training data to such an extent that it struggles to generalize well to unseen or new data.

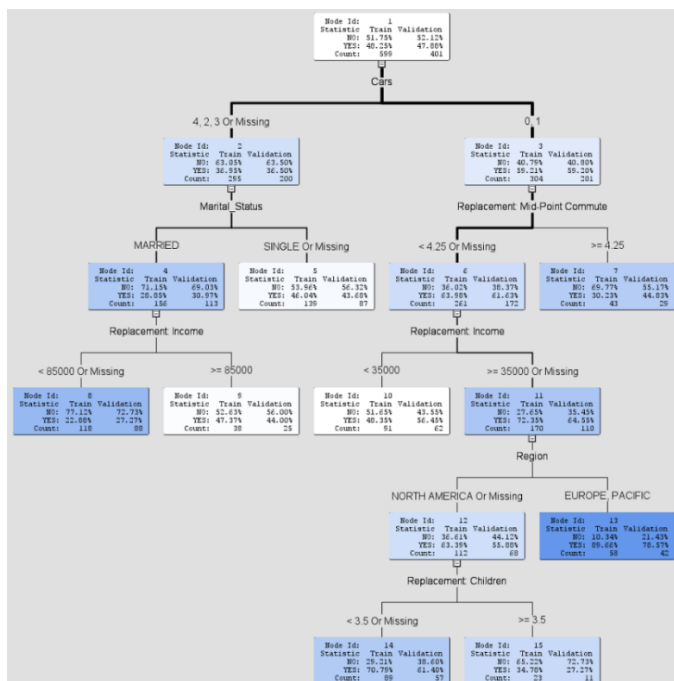
Below are the differences between both methods. Given our objectives and the need to balance complexity with performance, Method #1 emerged as the more suitable choice.

Method	Model Details	ASE Value
#1	Neural Network, 7 hidden units	0.215715
#2	Neural Network, 7 hidden units	0.222288

Method #1



Method #2



ASE values of Method #1, after running all models.

Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error
Neural9	Neural9	Neural Network 7H	Purchased...	Purchased ...	0.215715
Neural5	Neural5	Neural Network 3H	Purchased...	Purchased ...	0.215734
Neural10	Neural10	Neural Network 8H	Purchased...	Purchased ...	0.216029
Neural4	Neural4	Neural Network 2H	Purchased...	Purchased ...	0.222289
Neural7	Neural7	Neural Network 5H	Purchased...	Purchased ...	0.222486
Neural8	Neural8	Neural Network 6H	Purchased...	Purchased ...	0.226104
Neural6	Neural6	Neural Network 4H	Purchased...	Purchased ...	0.229208
Reg2	Reg2	Forward Regression	Purchased...	Purchased ...	0.232781
Reg4	Reg4	Stepwise Regression	Purchased...	Purchased ...	0.232781
Tree2	Tree2	ASE Decision Tree	Purchased...	Purchased ...	0.233013
Tree3	Tree3	Misclassification Decision Tree	Purchased...	Purchased ...	0.236706
Neural	Neural	Neural Network	Purchased...	Purchased ...	0.237697
Reg3	Reg3	Backward Regression	Purchased...	Purchased ...	0.237917
Tree	Tree	Maximal Decision Tree	Purchased...	Purchased ...	0.23915
Reg	Reg	Full Regression	Purchased...	Purchased ...	0.239344

The Method #1 yielded us an **ASE value of 0.215715** for the Neural Network model with **7 hidden units**. This was better than the model with 8 hidden units as well. We ended up selecting this model as it was the best compared to any other.

ASE values of Method #2, after running all models.

Predecessor Node	Model Node	Model Description	Valid: Average Squared Error ▲
Neural5	Neural5	Neural Network 7H	0.222288
Neural6	Neural6	Neural Network 8H	0.224099
Neural8	Neural8	Neural Network 9H	0.227121
Neural2	Neural2	Neural Network 5H	0.229655
Neural3	Neural3	Neural Network 6H	0.229666
Tree	Tree	Maximal Tree	0.230784
Tree3	Tree3	ASE Tree	0.230784
Neural4	Neural4	Neural Network 4H	0.23197
Neural	Neural	Neural Network 3H	0.232784
Reg2	Reg2	Stepwise Regression	0.232984
Reg4	Reg4	Forward Regression	0.232984
Neural7	Neural7	Neural Network 2H	0.233697
Reg3	Reg3	Backward Regression	0.236121
Reg	Reg	Regression	0.236622
Tree2	Tree2	Misclassification Tree	0.236706

Method #2's Neural Network model with 7 hidden units also gave us the project's lowest ASE value, at 0.222288. However comparing ASE values of method 1 & 2, we naturally selected #1 due to **the lower ASE value**.

Replace/Filter

The Replacement Node in SAS Enterprise Miner handles outliers or extreme values in interval variables. Here's a concise summary:

- Purpose: Addresses incorrect or extreme values in numeric variables.
- Location: Found under the "Modify" tab in SAS Enterprise Miner.
- Default Behavior: Identifies and replaces values more than three standard deviations from the mean.

Data Partition

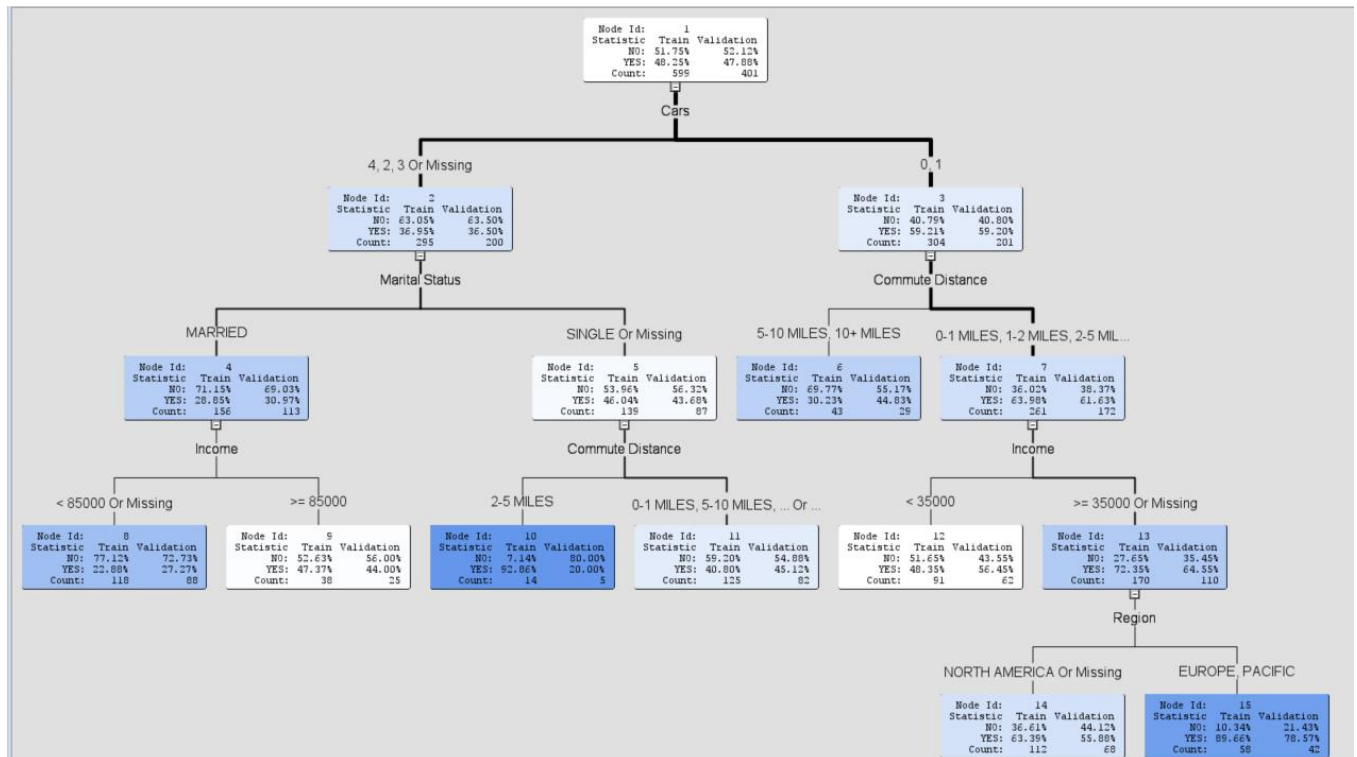
Property	Value
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	60.0
Validation	40.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	04/12/23 10:31 PM
Run ID	0784c12a-3050-4652-4
Last Error	
Last Status	Complete
Last Run Time	04/12/23 10:52 PM
Run Duration	0 Hr. 0 Min. 5.84 Sec.
Grid Host	
User-Added Node	No

Output			
1	-----		
2	User:	301333781	
3	Date:	December 04, 2023	
4	Time:	22:52:39	
5	-----		
6	* Training Output		
7	-----		
8			
9			
10			
11			
12	Variable Summary		
13			
14		Measurement	Frequency
15	Role	Level	Count
16			
17	ID	NOMINAL	1
18	INPUT	INTERVAL	2
19	INPUT	NOMINAL	8
20	INPUT	ORDINAL	1
21	TARGET	BINARY	1
22			
23			
24			
25			
26	Partition Summary		
27			
28			
29	Type	Data Set	Number of
30			Observations
31	DATA	EMWS1.FIMPORT_train	1000
32	TRAIN	EMWS1.Part_TRAIN	599
33	VALIDATE	EMWS1.Part_VALIDATE	401

Data Type	Description
Training Data	Used to train the model by allowing it to learn patterns, relationships, and structures within the dataset.
Validation Data	Used to fine-tune the model and assess its performance on unseen data, helping improve its generalization.
Test Data	Used to evaluate the final model's performance on completely new data, providing an unbiased measure of accuracy.

Data Partition	Percentage	Number of Observations
Training Data	60%	599
Validation Data	40%	401
Test Data	0%	0

Maximal Tree



A maximal tree is a fully expanded decision tree grown without any constraints or pruning. It aims to capture all possible details and complexities from the training data, which may lead to overfitting.

1. Complexity

- **Description:** Maximal trees are highly complex with numerous nodes and branches.
- **Detail:** They encompass all possible splits based on the available predictors, capturing even minor patterns and noise in the training data.

2. Overfitting

- **Risk:** Due to their high complexity, maximal trees are prone to overfitting.
- **Impact:** While they may perfectly fit the training data, they often struggle to generalize well to unseen or new data.

3. High Variance

- **Issue:** The variance of predictions tends to be high.
- **Reason:** The tree captures noise and specificities of the training data, making it less effective when applied to new data.

Method and Assessment

Method (Largest)

- **Description:** This setting instructs the algorithm to grow the tree to its maximum complexity.
- **Detail:** The tree is expanded to its largest size without any constraints.

Assessment Measure (Decision)

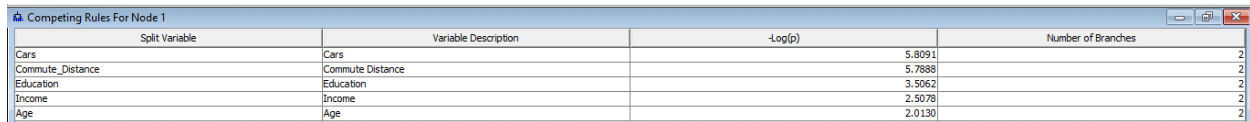
- **Description:** Uses a decision-based assessment measure to determine the splits and branches.
- **Detail:** Optimizes the tree structure based on certain decision criteria.

Default Settings

- **Configuration:** Everything else was set to the default settings.
- **Outcome:** After growing the tree to its fullest extent, it comprised 8 terminal nodes or leaves.

Seeing 8 leaves in the maximal tree indicates that after growing the tree to its fullest extent, it's comprised of 8 terminal nodes or leaves. Each leaf represents a final decision or classification made by the tree based on the combinations of predictor variables

Competing Splits for Maximal Tree



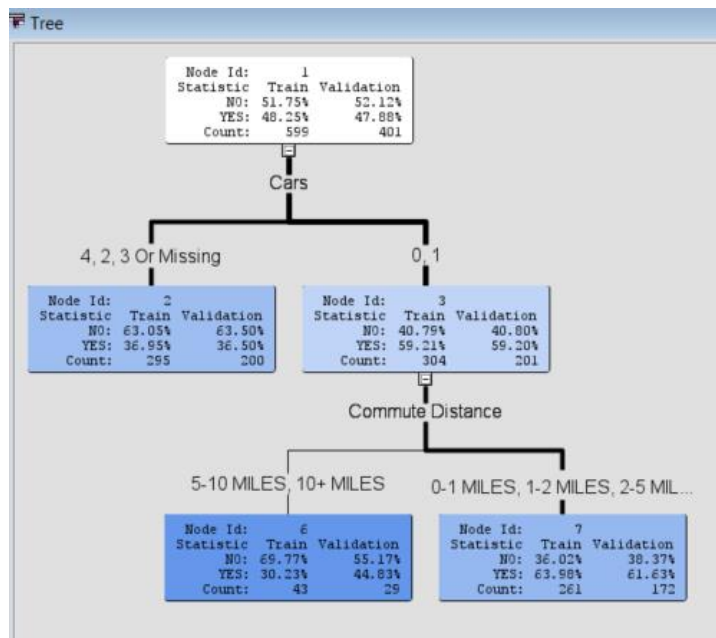
Split Variable	Variable Description	-Log(p)	Number of Branches
Cars	Cars	5.8091	2
Commute_Distance	Commute Distance	5.7888	2
Education	Education	3.5062	2
Income	Income	2.5078	2
Age	Age	2.0130	2

The competing splits screenshot shows us the variables from which the splits have been made. It has been sorted by best to worst. Our best variable for the split was **Cars**. The competing variables for the splits were **Commute_Distance, Education and Income**.

Misclassification Tree

A misclassification tree is a type of decision tree that prioritizes reducing misclassification errors during the tree-building process. It specifically uses splitting criteria that aim to minimize the misclassification rate when creating decision rules.

Feature	Description
Objective	Minimize misclassification rates, unlike other trees using information gain or Gini impurity.
Method (Assessment)	Refers to the method used for assessing the tree's performance during its construction.
Assessment Measure	Misclassification, indicating focus on reducing misclassification errors at each node.



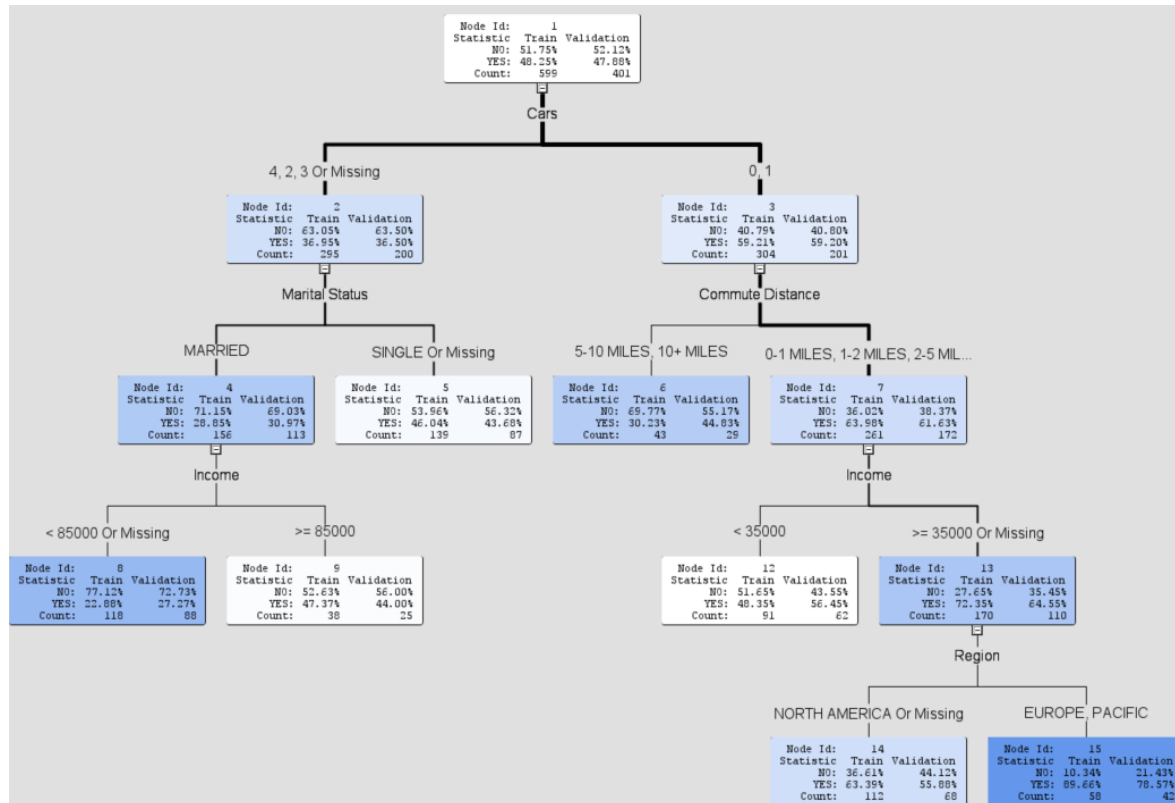
Competing Splits for Misclassification Tree

Competing Rules For Node 1				
Split Variable	Variable Description	-Log(p)	Number of Branches	
Cars	Cars	5.8091	2	
Commute_Distance	Commute Distance	5.7888	2	
Education	Education	3.5062	2	
Income	Income	2.5078	2	
Age	Age	2.0130	2	
Region	Region	1.5084	2	
Marital_Status	Marital Status	1.3435	2	
Children	Children	0.9587	2	
Occupation	Occupation	0.8936	2	
Home_Owner	Home Owner	0.3602	2	
Gender	Gender	0.1767	2	

Our best variable for the split was **Cars**. The competing variables for the splits were **Commute_Distance**, **Education** and **Income**.

Average Square Error Tree

An Average Square Error (ASE) tree aims to minimize the overall error by considering various variable subsets at each node. Unlike classification trees that focus on minimizing misclassification rates, ASE trees are used for regression problems, aiming to reduce the variance in predicted outcomes by considering different combinations of variables.



Feature	Description
Objective	Minimize the average square error (ASE), focusing on reducing the overall prediction error.
Method (Assessment)	Uses ASE to evaluate the performance of the tree during its construction, ensuring low error rates.
Assessment Measure	Average Square Error, indicating the focus on minimizing the squared differences between predicted and actual values at each node.

Competing Splits for ASE Tree

Competing Rules For Node 1			
Split Variable	Variable Description	-Log(p)	Number of Branches
Cars	Cars	5.8091	2
Commute_Distance	Commute Distance	5.7888	2
Education	Education	3.5062	2
Income	Income	2.5078	2
Age	Age	2.0130	2
Region	Region	1.5084	2
Marital_Status	Marital Status	1.3435	2
Children	Children	0.9587	2
Occupation	Occupation	0.8936	2
Home_Owner	Home Owner	0.3602	2
Gender	Gender	0.1767	2

Our analysis of competing splits revealed important factors influencing bike purchases. The number of **cars** possessed consistently emerged as the most significant predictor, followed by commute distance, education level, and income. This hierarchy of importance provides valuable insights for targeted marketing strategies. Businesses could focus their efforts on consumers based on car ownership status and commuting habits, potentially increasing the effectiveness of their campaigns.

Variable	Importance
Cars	High
Commute Distance	Medium
Education	Medium
Income	Low
Age	Low
Region	Low
Marital Status	Low
Children	Low
Occupation	Low
Home Owner	Low
Gender	Low

Impute

Variable Name	Impute Method	Imputed Variable	Indicator Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Age	MEAN	IMP Age	M Age	43.877104377	INPUT	INTERVAL		5
Cars	COUNT	IMP Cars	M Cars	2	INPUT	NOMINAL		7
Children	COUNT	IMP Children	M Children	0	INPUT	NOMINAL		7
Gender	COUNT	IMP Gender	M Gender	Male	INPUT	NOMINAL		9
Home Owner	COUNT	IMP Home Owner	M Home Owner	Yes	INPUT	NOMINAL	Home Owner	3
Income	MEAN	IMP Income	M Income	56275.167785	INPUT	INTERVAL		3
Marital Status	COUNT	IMP Marital Status	M Marital Status	Married	INPUT	NOMINAL	Marital Status	5

Variables without Missing Values:

Variable	Description
ID	Unique identification numbers, complete without missing entries.
Purchased Bike	Indicates bike purchase, also without missing values.
Region	Categorical variable denoting geographic regions.
Commute Distance	Represents commuting distance, complete without missing data.
Education	Indicates the level of education attained by individuals.
Occupation	Specifies the profession or type of occupation.

Variables with Missing Values (Filled using Imputation):

Variable	Description
Age	Information about individuals' ages, missing values have been filled.
Cars	Indicates the number of cars owned, missing values have been imputed.
Children	Information about the number of children an individual has, missing values imputed.
Gender	Specifies the gender of individuals, missing values filled using imputation.
Home Owner	Indicates whether individuals own a home, missing values have been imputed.
Income	Represents individuals' income levels, missing values have been imputed.
Marital Status	Information about the marital status of individuals, missing values have been

Imputation involves estimating or predicting missing values based on the available data. It allowed us to retain records with missing values by filling in those gaps, making the dataset more complete for subsequent analysis and modeling.

Skewness

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	Purchas...	No	IMP Inco...	50000	0	310	10000	170000	53096.77	30909.99	0.812879	0.713627	INPUT	Imputed I...	-0.05648	0.060584	1
TRAIN	Purchas...	Yes	IMP Inco...	60000	0	289	10000	170000	59684.52	31820.51	0.937363	0.971795	INPUT	Imputed I...	0.060584	0.060584	2
TRAIN	Purchas...	No	IMP Age	44	0	310	25	78	44.84437	12.0979	0.263359	-0.79476	INPUT	Imputed ...	0.022045	0.023647	1
TRAIN	Purchas...	Yes	IMP Age	41	0	289	25	78	42.83955	9.717515	0.688689	0.365398	INPUT	Imputed ...	-0.02365	0.023647	2

Skewness measures a variable's asymmetry in distribution. Skewness scores below one, with the greatest at 0.937, imply that our variables' distributions are relatively balanced. As a result of the lack of skewness, we did not cap and floor or transform the variables further. This discovery is important for businesses since it shows that our dataset was well-suited to standard modelling techniques without the need for complex modification.

Variable	Skewness Value
Income	0.937
Age	0.812

Full Regression

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Purchased_Bike	Purchased Bike	_AIC_	Akaike's Information Criterion	762.677		
Purchased_Bike	Purchased Bike	_ASE_	Average Squared Error	0.200433		0.239344
Purchased_Bike	Purchased Bike	_AVERR_	Average Error Function	0.588211		0.683962
Purchased_Bike	Purchased Bike	_DFE_	Degrees of Freedom for Error	570		
Purchased_Bike	Purchased Bike	_DFM_	Model Degrees of Freedom	29		
Purchased_Bike	Purchased Bike	_DFT_	Total Degrees of Freedom	599		
Purchased_Bike	Purchased Bike	_DIV_	Divisor for ASE	1198		802
Purchased_Bike	Purchased Bike	_ERR_	Error Function	704.677		548.5374
Purchased_Bike	Purchased Bike	_FPE_	Final Prediction Error	0.220828		
Purchased_Bike	Purchased Bike	_MAX_	Maximum Absolute Error	0.957646		0.966224
Purchased_Bike	Purchased Bike	_MSE_	Mean Square Error	0.210631		0.239344
Purchased_Bike	Purchased Bike	_NOBS_	Sum of Frequencies	599		401
Purchased_Bike	Purchased Bike	_NW_	Number of Estimate Weights	29		
Purchased_Bike	Purchased Bike	_RASE_	Root Average Sum of Squares	0.447698		0.489228
Purchased_Bike	Purchased Bike	_RFPE_	Root Final Prediction Error	0.469924		
Purchased_Bike	Purchased Bike	_RMSE_	Root Mean Squared Error	0.458945		0.489228
Purchased_Bike	Purchased Bike	_SBC_	Schwarz's Bayesian Criterion	890.1396		
Purchased_Bike	Purchased Bike	_SSE_	Sum of Squared Errors	240.1191		191.9539
Purchased_Bike	Purchased Bike	_SUMW_	Sum of Case Weights Times Freq	1198		802
Purchased_Bike	Purchased Bike	_MISC_	Misclassification Rate	0.315526		0.391521

The full regression model evaluated all available factors to predict bike purchases. While comprehensive, this strategy may not always be the most efficient or effective for firms. It serves as a baseline for comparison with more sophisticated models and can identify possible areas of interest that might otherwise be overlooked in more selective approaches.

Forward Regression

Fit Statistics					
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Purchased_Bike	Purchased Bike	_AIC_	Akaike's Information Criterion	760.9749	
Purchased_Bike	Purchased Bike	_ASE_	Average Squared Error	0.207698	0.232781
Purchased_Bike	Purchased Bike	_AVERR_	Average Error Function	0.605154	0.66484
Purchased_Bike	Purchased Bike	_DFE_	Degrees of Freedom for Error	581	
Purchased_Bike	Purchased Bike	_DFM_	Model Degrees of Freedom	18	
Purchased_Bike	Purchased Bike	_DFT_	Total Degrees of Freedom	599	
Purchased_Bike	Purchased Bike	_DIV_	Divisor for ASE	1198	802
Purchased_Bike	Purchased Bike	_ERR_	Error Function	724.9749	533.2016
Purchased_Bike	Purchased Bike	_FPE_	Final Prediction Error	0.220568	
Purchased_Bike	Purchased Bike	_MAX_	Maximum Absolute Error	0.965952	0.954399
Purchased_Bike	Purchased Bike	_MSE_	Mean Square Error	0.214133	0.232781
Purchased_Bike	Purchased Bike	_NOBS_	Sum of Frequencies	599	401
Purchased_Bike	Purchased Bike	_NW_	Number of Estimate Weights	18	
Purchased_Bike	Purchased Bike	_RASE_	Root Average Sum of Squares	0.455739	0.482474
Purchased_Bike	Purchased Bike	_RFPE_	Root Final Prediction Error	0.469646	
Purchased_Bike	Purchased Bike	_RMSE_	Root Mean Squared Error	0.462745	0.482474
Purchased_Bike	Purchased Bike	_SBC_	Schwarz's Bayesian Criterion	840.0896	
Purchased_Bike	Purchased Bike	_SSE_	Sum of Squared Errors	248.8226	186.6905
Purchased_Bike	Purchased Bike	_SUMW_	Sum of Case Weights Times Freq	1198	802
Purchased_Bike	Purchased Bike	_MISC_	Misclassification Rate	0.308848	0.376559

Our forward regression model was the best predictor of bike purchases, with an average squared error of 0.232781. This model gradually added the most important factors, culminating in a streamlined but effective forecasting tool. This technique strikes a compromise between complexity and accuracy, which could lead to more actionable insights and efficient resource allocation in marketing and sales strategies.

Backward Regression

Fit Statistics					
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Purchased_Bike	Purchased Bike	_AIC_	Akaike's Information Criterion	757.8624	
Purchased_Bike	Purchased Bike	_ASE_	Average Squared Error	0.203509	0.237917
Purchased_Bike	Purchased Bike	_AVERR_	Average Error Function	0.595878	0.676383
Purchased_Bike	Purchased Bike	_DFE_	Degrees of Freedom for Error	577	
Purchased_Bike	Purchased Bike	_DFM_	Model Degrees of Freedom	22	
Purchased_Bike	Purchased Bike	_DFT_	Total Degrees of Freedom	599	
Purchased_Bike	Purchased Bike	_DIV_	Divisor for ASE	1198	802
Purchased_Bike	Purchased Bike	_ERR_	Error Function	713.8624	542.4593
Purchased_Bike	Purchased Bike	_FPE_	Final Prediction Error	0.219028	
Purchased_Bike	Purchased Bike	_MAX_	Maximum Absolute Error	0.952507	0.953082
Purchased_Bike	Purchased Bike	_MSE_	Mean Square Error	0.211268	0.237917
Purchased_Bike	Purchased Bike	_NOBS_	Sum of Frequencies	599	401
Purchased_Bike	Purchased Bike	_NW_	Number of Estimate Weights	22	
Purchased_Bike	Purchased Bike	_RASE_	Root Average Sum of Squares	0.451119	0.487767
Purchased_Bike	Purchased Bike	_RFPE_	Root Final Prediction Error	0.468004	
Purchased_Bike	Purchased Bike	_RMSE_	Root Mean Squared Error	0.459639	0.487767
Purchased_Bike	Purchased Bike	_SBC_	Schwarz's Bayesian Criterion	854.5581	
Purchased_Bike	Purchased Bike	_SSE_	Sum of Squared Errors	243.8034	190.8093
Purchased_Bike	Purchased Bike	_SUMW_	Sum of Case Weights Times Freq	1198	802
Purchased_Bike	Purchased Bike	_MISC_	Misclassification Rate	0.308848	0.386534

The backward regression model began with all variables and gradually excluded less significant ones, resulting in an average squared error of 0.237917. While not as effective as forward regression in this scenario, this method can help discover which elements are less important in predicting customer behavior, thereby saving resources by focusing on the most relevant variables.

Stepwise Regression

Fit Statistics					
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Purchased_Bike	Purchased Bike	_AIC_	Akaike's Information Criterion	760.9749	.
Purchased_Bike	Purchased Bike	_ASE_	Average Squared Error	0.207698	0.232781
Purchased_Bike	Purchased Bike	_AVER_	Average Error Function	0.605154	0.66484
Purchased_Bike	Purchased Bike	_DFE_	Degrees of Freedom for Error	581	.
Purchased_Bike	Purchased Bike	_DFM_	Model Degrees of Freedom	18	.
Purchased_Bike	Purchased Bike	_DFT_	Total Degrees of Freedom	599	.
Purchased_Bike	Purchased Bike	_DIV_	Divisor for ASE	1198	802
Purchased_Bike	Purchased Bike	_ERR_	Error Function	724.9749	533.2016
Purchased_Bike	Purchased Bike	_FPE_	Final Prediction Error	0.220568	.
Purchased_Bike	Purchased Bike	_MAX_	Maximum Absolute Error	0.965952	0.954399
Purchased_Bike	Purchased Bike	_MSE_	Mean Square Error	0.214133	0.232781
Purchased_Bike	Purchased Bike	_NOBS_	Sum of Frequencies	599	401
Purchased_Bike	Purchased Bike	_NW_	Number of Estimate Weights	18	.
Purchased_Bike	Purchased Bike	_RASE_	Root Average Sum of Squares	0.455739	0.482474
Purchased_Bike	Purchased Bike	_RFPE_	Root Final Prediction Error	0.469646	.
Purchased_Bike	Purchased Bike	_RMSE_	Root Mean Squared Error	0.462745	0.482474
Purchased_Bike	Purchased Bike	_SBC_	Schwarz's Bayesian Criterion	840.0896	.
Purchased_Bike	Purchased Bike	_SSE_	Sum of Squared Errors	248.8226	186.6905
Purchased_Bike	Purchased Bike	_SUMW_	Sum of Case Weights Times Freq	1198	802
Purchased_Bike	Purchased Bike	_MISC_	Misclassification Rate	0.308848	0.376559

We performed a stepwise regression, which combines forward and backward techniques. It's another approach to support our findings. While it did not exceed our forward regression, it increased our confidence in the results.

Regression Analysis

Regression Model	ASE Value
Full Regression	0.239344
Forward Regression	0.232781
Backward Regression	0.237917
Stepwise Regression	0.232781

Insight: Forward Regression has the lowest ASE value, making it the best predictor among regression models.

Interpretation of Regression (forward)

Results - Node: Forward Regression Diagram: Bike Buyers

File Edit View Window

Output

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp (Est)
Intercept	1	-1.0674	0.3719	8.24	0.0041		0.344
Commute_Distance 0-1 Miles	1	0.5861	0.1865	9.88	0.0017		1.797
Commute_Distance 1-2 Miles	1	0.1678	0.2080	0.65	0.4199		1.183
Commute_Distance 10+ Miles	1	-0.9677	0.2921	10.98	0.0009		0.380
Commute_Distance 2-5 Miles	1	0.8096	0.2073	15.25	<.0001		2.247
IMP_Cars 0	1	0.5005	0.2310	4.70	0.0302		1.650
IMP_Cars 1	1	0.4376	0.2086	4.40	0.0359		1.549
IMP_Cars 2	1	-0.2361	0.2062	1.31	0.2522		0.790
IMP_Cars 3	1	-0.1151	0.3387	0.12	0.7340		0.891
IMP_Children 0	1	0.1794	0.1816	0.98	0.3232		1.196
IMP_Children 1	1	0.1932	0.2306	0.70	0.4020		1.213
IMP_Children 2	1	0.2042	0.2028	1.01	0.3140		1.227
IMP_Children 3	1	0.4198	0.2329	3.25	0.0715		1.522
IMP_Children 4	1	0.1663	0.2280	0.53	0.4656		1.181
IMP_Income	1	0.000012	5.609E-6	4.81	0.0283	0.2136	1.000
IMP_Marital_Status Married	1	-0.2815	0.0986	8.16	0.0043		0.755
Occupation Clerical	1	-0.2029	0.2431	0.70	0.4039		0.816
Occupation Management	1	-0.1513	0.2990	0.26	0.6130		0.860
Occupation Manual	1	-0.5024	0.3452	2.12	0.1456		0.605
Occupation Professional	1	0.6495	0.2378	7.46	0.0063		1.915
Region Europe	1	-0.0774	0.1789	0.19	0.6651		0.925
Region North America	1	-0.5557	0.1548	12.89	0.0003		0.574

Odds Ratio Estimates

Results - Node: Forward Regression Diagram: Bike Buyers

File Edit View Window

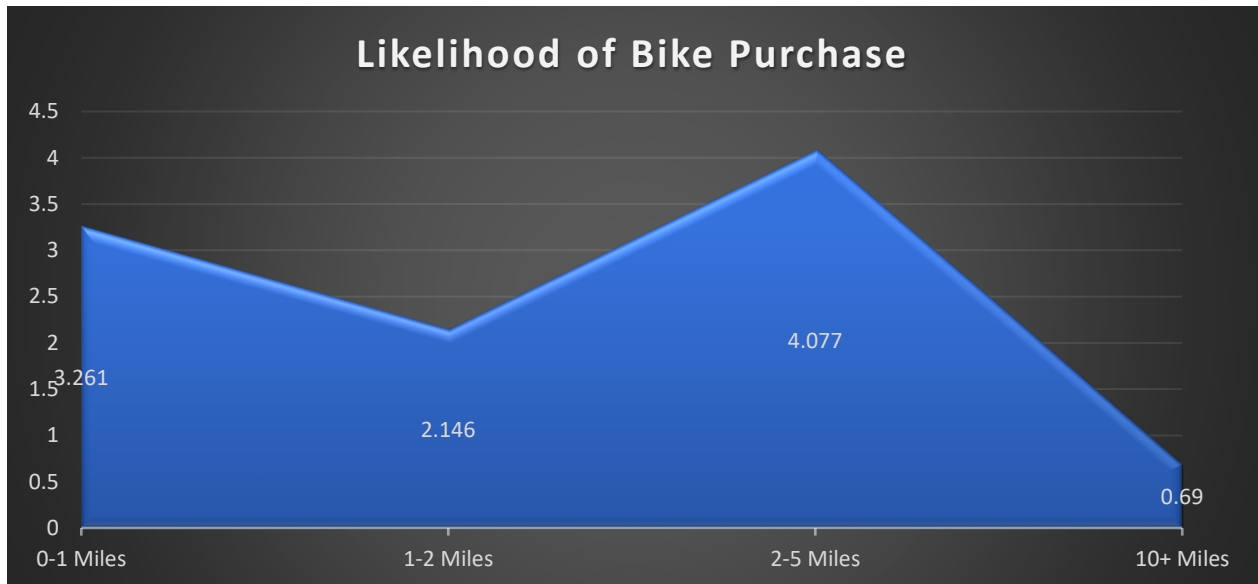
Output

Effect	Point Estimate
Commute_Distance 0-1 Miles vs 5-10 Miles	3.261
Commute_Distance 1-2 Miles vs 5-10 Miles	2.146
Commute_Distance 10+ Miles vs 5-10 Miles	0.689
Commute_Distance 2-5 Miles vs 5-10 Miles	4.077
IMP_Cars 0 vs 4	2.967
IMP_Cars 1 vs 4	2.786
IMP_Cars 2 vs 4	1.420
IMP_Cars 3 vs 4	1.603
IMP_Children 0 vs 5	3.828
IMP_Children 1 vs 5	3.881
IMP_Children 2 vs 5	3.924
IMP_Children 3 vs 5	4.869
IMP_Children 4 vs 5	3.778
IMP_Income	1.000
IMP_Marital_Status Married vs Single	0.569
Occupation Clerical vs Skilled Manual	0.664
Occupation Management vs Skilled Manual	0.699
Occupation Manual vs Skilled Manual	0.492
Occupation Professional vs Skilled Manual	1.556
Region Europe vs Pacific	0.491
Region North America vs Pacific	0.305

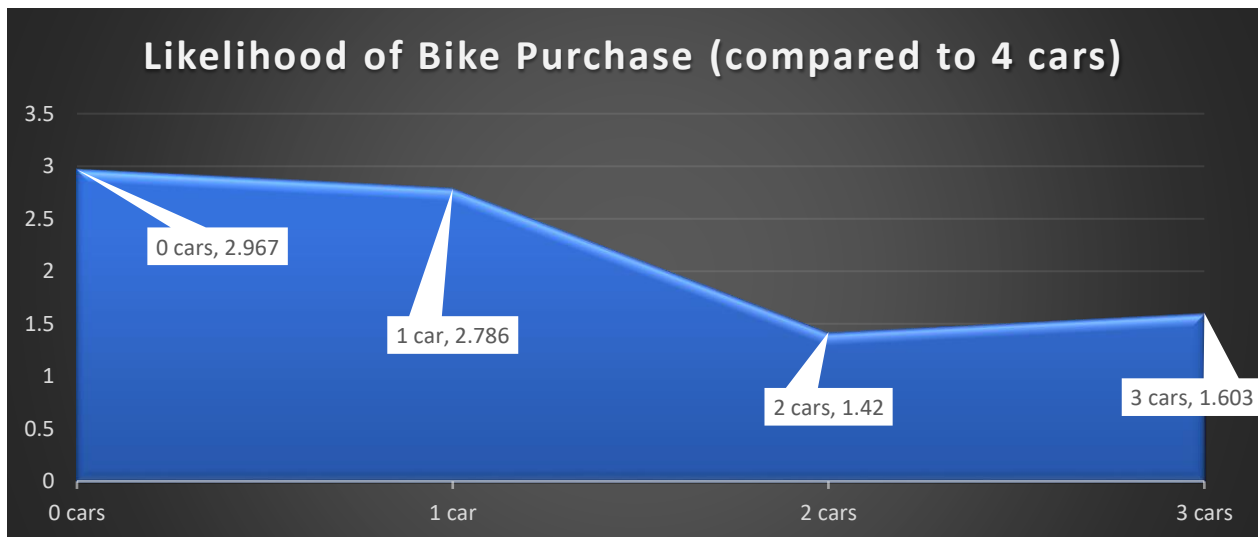
NOTE: No (additional) effects met the 0.05 significance level for entry into the model.

Interpretations of the Odds Ratio Estimates are as follows:

1. Commute Distance: Compared to 5-10 miles

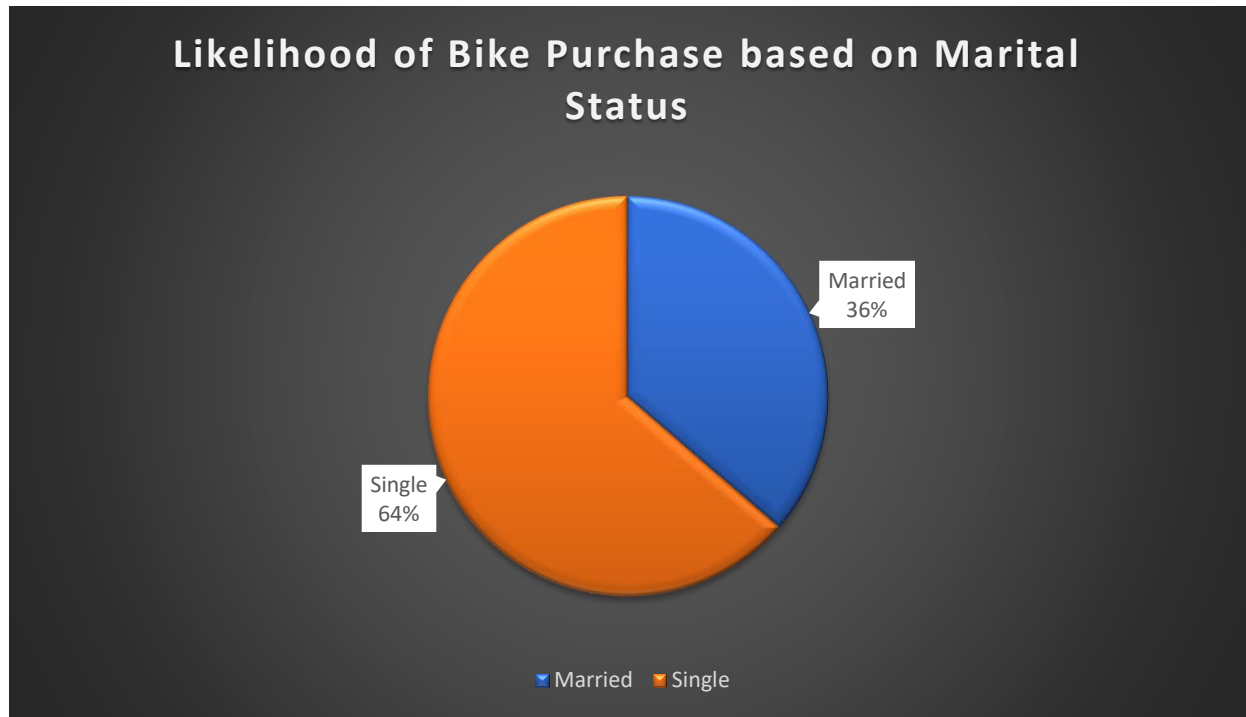


2. IMP_Cars (On the basis of Cars): Compared to 4 cars



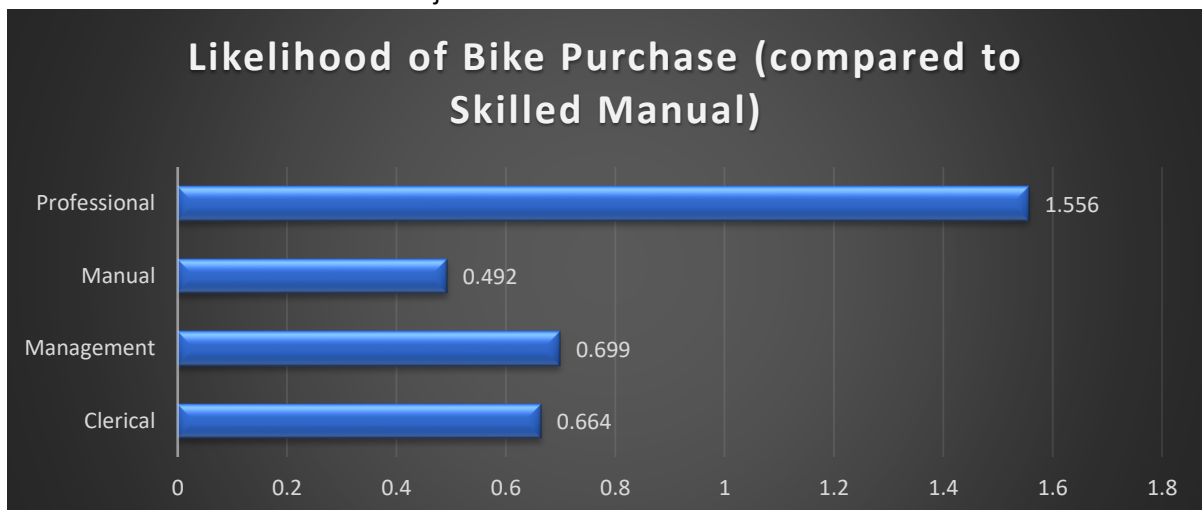
3. IMP_Income (On the basis of Income): There is no change based on the income of an individual to the likeliness of them purchasing a bike.

4. IMP_Marital_Status (On the basis of Marital Status): Married individuals are 43% less likely to purchase a bike than single individuals.



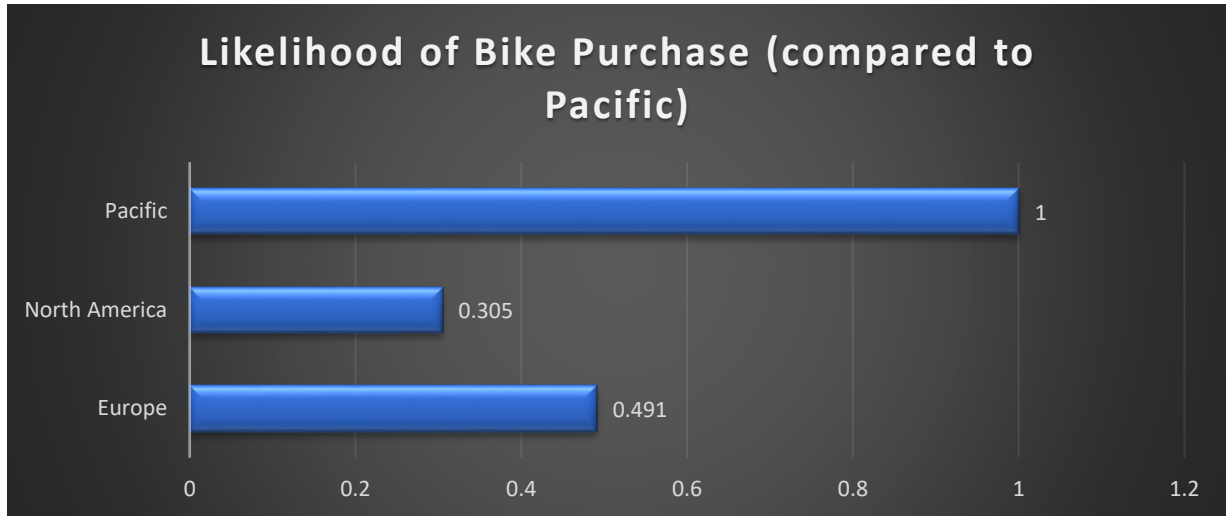
5. Occupation: Compared to 'Skilled Manual' Occupation:

- Clerical- An individual who has a clerical job is 33.6% less likely to purchase a bike than an individual who has a skilled manual job.
- Management- An individual who has a management job is 30.1% less likely to purchase a bike than an individual who has a skilled manual job.
- Manual- An individual who has a manual job is 50.8% less likely to purchase a bike than an individual who has a skilled manual job.
- Professional- An individual who has a professional job is 55.6% more likely to purchase a bike than an individual who has a skilled manual job.



6. Region: Compared to an individual living in the Pacific region:

- Europe- A person living in Europe is 50.9% less likely to purchase a bike than a person living in the Pacific region.
- North America- A person living in North America is 69.5% less likely to purchase a bike than a person living in the Pacific Region.



Full Neural Network (from impute)

Results - Node: Neural Network: Diagram: Bike Buyers

File Edit View Window

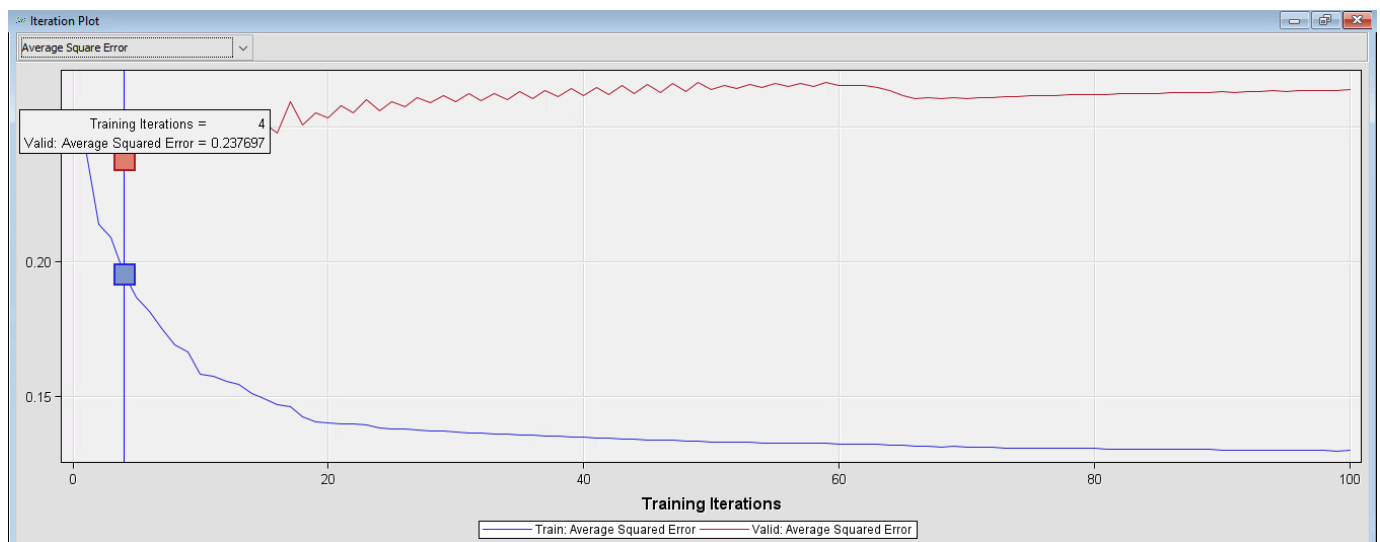
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Purchased_Bike	Purchased Bike	_DFT_	Total Degrees of Freedom	599		
Purchased_Bike	Purchased Bike	_DFE_	Degrees of Freedom for Error	508		
Purchased_Bike	Purchased Bike	_DFM_	Model Degrees of Freedom	91		
Purchased_Bike	Purchased Bike	_NW_	Number of Estimated Weights	91		
Purchased_Bike	Purchased Bike	_AIC_	Akaike's Information Criterion	865.2319		
Purchased_Bike	Purchased Bike	_SBC_	Schwarz's Bayesian Criterion	1265.201		
Purchased_Bike	Purchased Bike	_ASE_	Average Squared Error	0.195278		0.237697
Purchased_Bike	Purchased Bike	_MAX_	Maximum Absolute Error	0.895031		0.977622
Purchased_Bike	Purchased Bike	_DIV_	Divisor for ASE	1198		802
Purchased_Bike	Purchased Bike	_NOBS_	Sum of Frequencies	599		401
Purchased_Bike	Purchased Bike	_RASE_	Root Average Squared Error	0.441903		0.487542
Purchased_Bike	Purchased Bike	_SSE_	Sum of Squared Errors	233.9429		190.6331
Purchased_Bike	Purchased Bike	_SUMW_	Sum of Case Weights Times Freq	1198		802
Purchased_Bike	Purchased Bike	_FPE_	Final Prediction Error	0.26524		
Purchased_Bike	Purchased Bike	_MSE_	Mean Squared Error	0.230259		0.237697
Purchased_Bike	Purchased Bike	_RFPE_	Root Final Prediction Error	0.515014		
Purchased_Bike	Purchased Bike	_RMSE_	Root Mean Squared Error	0.479853		0.487542
Purchased_Bike	Purchased Bike	_AVER_	Average Error Function	0.57031		0.679419
Purchased_Bike	Purchased Bike	_ERR_	Error Function	683.2319		544.8939
Purchased_Bike	Purchased Bike	_MISC_	Misclassification Rate	0.308848		0.394015
Purchased_Bike	Purchased Bike	_WRONG_	Number of Wrong Classifications	185		158

Iteration Plot

We analyzed the performance of several neural networks to find the best one for our needs. One of these networks, called the Impute Neural Network, was tested over a range of up to 100 iterations. Our analysis showed that this Impute Neural Network did not fully stabilize within those 100 cycles. Interestingly, the lowest average error, a measure of how accurately the network predicts, was found at the fourth cycle, with a value of 0.237697.

Among all the networks we tested, the best performer was the one derived from a method called Forward Regression. This network, which had seven hidden units (think of these as internal nodes that help process the information), achieved the lowest error rate of 0.215715. This means it made the most accurate predictions compared to the others.



Optimal Neural Network (forward)

Results - Node: Neural Network 7H Diagram: Bike Buyers

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Purchased_Bike	Purchased Bike	_DFT_	Total Degrees of Freedom	599		
Purchased_Bike	Purchased Bike	_DFE_	Degrees of Freedom for Error	465		
Purchased_Bike	Purchased Bike	_DFM_	Model Degrees of Freedom	134		
Purchased_Bike	Purchased Bike	_NW_	Number of Estimated Weights	134		
Purchased_Bike	Purchased Bike	_AIC_	Akaike's Information Criterion	934.5291		
Purchased_Bike	Purchased Bike	_SBC_	Schwarz's Bayesian Criterion	1523.494		
Purchased_Bike	Purchased Bike	_ASE_	Average Squared Error	0.18861		0.215715
Purchased_Bike	Purchased Bike	_MAX_	Maximum Absolute Error	0.934636		0.948729
Purchased_Bike	Purchased Bike	_DIV_	Divisor for ASE	1198		802
Purchased_Bike	Purchased Bike	_NOBS_	Sum of Frequencies	599		401
Purchased_Bike	Purchased Bike	_RASE_	Root Average Squared Error	0.434293		0.464451
Purchased_Bike	Purchased Bike	_SSE_	Sum of Squared Errors	225.9549		173.0031
Purchased_Bike	Purchased Bike	_SUMW_	Sum of Case Weights Times Freq	1198		802
Purchased_Bike	Purchased Bike	_FPE_	Final Prediction Error	0.297314		
Purchased_Bike	Purchased Bike	_MSE_	Mean Squared Error	0.242962		0.215715
Purchased_Bike	Purchased Bike	_RFPE_	Root Final Prediction Error	0.545266		
Purchased_Bike	Purchased Bike	_RMSE_	Root Mean Squared Error	0.492912		0.464451
Purchased_Bike	Purchased Bike	_AVERR_	Average Error Function	0.556368		0.622782
Purchased_Bike	Purchased Bike	_ERR_	Error Function	666.5291		499.4715
Purchased_Bike	Purchased Bike	_MISC_	Misclassification Rate	0.283806		0.331671
Purchased_Bike	Purchased Bike	_WRONG_	Number of Wrong Classifications	170		133

Iteration Plot

The Neural Network with 7 Hidden Units is the Neural Network that is connected to the Forward Regression node. Our testing involved running the network for up to 100 cycles, or iterations. The results showed that the network didn't fully stabilize within those 100 cycles. However, the lowest average error rate, which tells us how accurate the network's predictions are, was achieved at the eighth cycle, with a value of 0.215715.



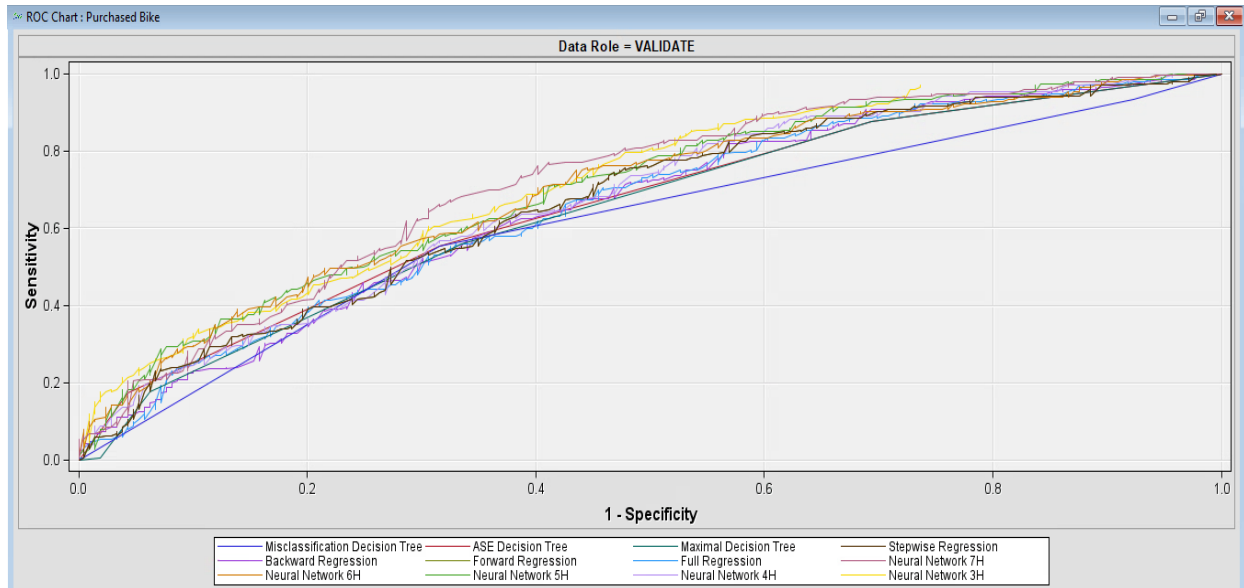
Assessment (model comparison)

We used the Model Comparison Node to evaluate and compare the performance of all significant models. This tool helps us identify which model is the most accurate and provides the best results.

Model Description	Average Squared Error	Target Variable
Neural Network 7H	0.215715	Purchased
Neural Network 3H	0.215734	Purchased
Neural Network 8H	0.216209	Purchased
Neural Network 2H	0.222289	Purchased
Neural Network 5H	0.222468	Purchased
Neural Network 6H	0.226104	Purchased
Forward Regression	0.227281	Purchased
Stepwise Regression	0.227831	Purchased
ASE Decision Tree	0.233701	Purchased
Misclassification Decision Tree	0.236706	Purchased
Neural Network	0.237697	Purchased
Backward Regression	0.239117	Purchased
Maximal Decision Tree	0.239315	Purchased
Full Regression	0.239344	Purchased

According to the comparison, the neural network with seven hidden units emerged as the most accurate model. It achieved an average error rate of 0.215715, making it the top performer in terms of prediction accuracy.

ROC Chart



An ROC Chart helps us understand the accuracy of a model by comparing correct and incorrect predictions at different points. It shows how well a model can make accurate predictions by measuring the area under its curve. The larger this area, the more accurate the model is.

Model Description	Average Squared Error	Valid ROC Index
Neural Network 7H	0.215715	0.716
Neural Network 3H	0.215734	0.718
Neural Network 8H	0.216209	0.733
Neural Network 2H	0.222289	0.701
Neural Network 5H	0.222468	0.702
Neural Network 6H	0.226104	0.674
Neural Network 4H	0.229208	0.649
Forward Regression	0.227281	0.669
Stepwise Regression	0.227831	0.669
ASE Decision Tree	0.233701	0.659

Misclassification Decision Tree	0.236706	0.658
Neural Network	0.237697	0.657
Backward Regression	0.239117	0.656
Maximal Decision Tree	0.239315	0.648
Full Regression	0.239344	0.658

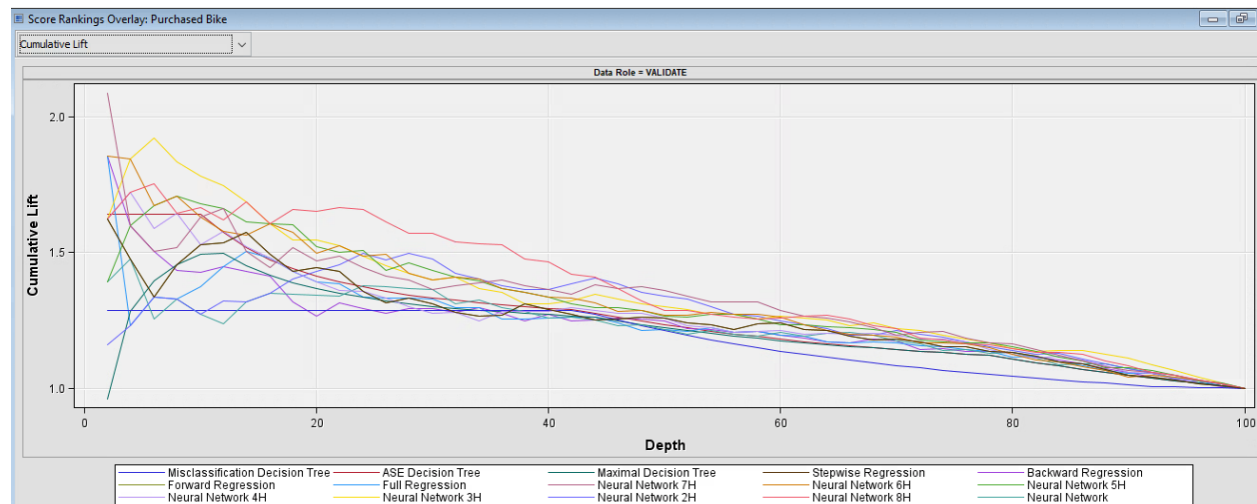
We use a value called the ROC Index to evaluate this. If the ROC Index is less than 0.6, the model is considered weak. If it's greater than 0.7, the model is considered strong.

The neural network with seven hidden units, which we selected as the best model, has a ROC Index of 0.716. This means it is a strong and accurate model, making it suitable for predicting Bike Buyers.

Lift Chart

A lift chart helps us understand how much more accurate our model is compared to making random guesses. It shows how much better our model's predictions are compared to just guessing without a model.

The main purpose of a lift chart is to evaluate the performance of our model against a baseline, which is random guessing. In simple terms, it tells us how much value our model adds in making accurate predictions.



Using a lift chart, we can assess how effective our model is. When the lift value is high at certain points, it indicates that our model is much better than random guessing at identifying the target. This shows that the model is performing well in making accurate predictions.

Conclusion

Understanding why consumers buy bikes involves exploring various factors like income, location, education, and more. Our study used internal data to analyze how these factors influence bike purchases. We used different models to predict buying behavior and found that commute distance, car ownership, job type, and where someone lives significantly impact whether they'll buy a bike.

Our study showed that tweaking how we analyze data affects our predictions. Even small changes in how we categorize things or select methods can change the accuracy of our models. For instance, when we compared different ways of looking at commute distance, we found it really influences bike-buying decisions.

We also looked at variables like income, marital status, and occupation. While income didn't seem to affect bike purchases, being married or the kind of job someone has did make a difference. For example, people with certain jobs were more or less likely to buy bikes compared to others.

Overall, our study gives a clearer picture of why people buy bikes. It shows that where someone lives, what they do for a living, and how they commute can play a big role. Understanding these factors can help businesses and policymakers make smarter decisions in the bike market.

References

Statista. (n.d.). *Bicycles - Worldwide*. Retrieved July 9, 2024, from <https://www.statista.com/outlook/mmo/bicycles/worldwide>