EMBRY-RIDDLE Aeronautical University SCHOLARLY COMMONS

International Journal of Aviation, Aeronautics, and Aerospace

Volume 6 | Issue 5 Article 14

2019

Predicting the U.S. Airline Operating Profitability using Machine Learning Algorithms

WooJin Choi
Embry-Riddle Aeronautical University, choiw1@my.erau.edu
Mary B. O'Connor
Embry-Riddle Aeronautical University, oconnom6@my.erau.edu
Dothang Truong
DB/School of Graduate Studies, truongd@erau.edu

Introduction

Since the U.S. deregulation of airlines in 1978, there has been extreme competition among the U.S. airlines for market share and, eventually, profitability (An, Chen, Park, & Subrahmanian, 2017). Under continuous liberalization and the expansion of the low-cost component of the global airline industry, and the volatility of fuel prices, airlines' revenue and cost structure have become critical factors in yielding operating margins for airlines (Cronrath, 2017).

Although growth in demand, increased efficiency, and reduced interest expenses helped the airline industry improve profitability in 2016 (International Air Transport Association, 2017), the industry faces longer-term challenges to financial performance with rising costs, fluctuating oil prices, and increasing cost competition. According to Cronrath (2017), while the aggregate net profit of the airline industry was almost zero between 1970 and 2010, the amplitude of the profit cycle has been increasing over time. The cycling period has become shorter and deeper after 2007, which calls for a careful approach in airline management decisions (Cronrath, 2017). Because the airline industry operates on small profit margins, profitability of the airline industry is sensitive to internal and external elements of operation, such as fuel, labor force, maintenance, passenger and cargo business, and other ancillary businesses (Miranda, 2015).

Statement of the Problem

It is of critical importance for the airline industry to comprehend the complex relationship among operations revenue, expenses, and profitability to tackle the challenges both with managerial processes and changing external environments that may influence profitability (Scotti & Volta, 2017). Although there has been considerable work completed analyzing the financial performance of the airline industry, the prediction of airline profitability is not adequately precise and the research conducted in the analysis of both operations revenue and expenses for predicting or improving airline profitability has been limited.

The majority of previous research traditionally sought to identify factors that influence airline costs and productivity or has selected just a few key cost and revenue elements to determine a potential relationship with profitability (An et al., 2017). Therefore, consideration of both airline revenue and cost profile in an integrative approach is of critical importance to examine the airline financial structure and profitability.

Purpose Statement

Given the sophisticated financial situations within the airline industry, understanding the combined effect of operating revenue and cost on airline profitability is imperative. Our intent is to fill the gap in the literature regarding airline performance by providing insight into the impact of financial structure on

profitability of the airline industry. Using the accumulation of financial data generated by the U.S. airlines, data-mining tools can help achieve this goal. The purpose of the present study is to assess the impact of operating revenues and operating expenses on profitability of the U.S. major airlines and to develop a predictive model for airlines' profit and loss structure.

Research Question

This project sought to answer three research questions by developing predictive research models. The questions are as follows:

RQ1: What is the most influential factor among airline operating revenues and expenses as reported in data from the U.S. Bureau of Transportation Statistics (BTS) in predicting an airline's profitability, using monetary and percentage values as two scales of measure of input variables?

RQ2: Between the decision tree and regression models, which provides the best model for predicting the profitability of a major U.S. airline, using monetary and percentage values as two scales of measure of input variables?

RQ3: Between a monetary and percentage scale of input variables, which scale can better predict airline profitability? The term monetary and percentage values are defined in the methodology section of this paper.

Significance of the Study

In consideration of the increasing fluctuation in financial performance of the airline industry, the practical contribution of this study will determine the factors which can best predict profitability and assist airline management in making appropriate decisions to improve the financial structure of the airline. While most of the previous studies focused on key financial or operational factors in a selective approach, this study strives to examine the effect of all factors related to airport operation costs and revenues on airline profitability, and presents a profitability predictive model using decision tree and regression methods.

Assumptions and Delimitations

This study focuses on the major U.S. commercial airlines with annual operating revenues of $1 billion or more per airline. In this study, the primary focus is on the major airlines' operating profit and loss distribution for the last decade, while fixed costs and non-operating income and expenses such as interest, taxes, and capital gains were not considered. Also, changes in the cost and revenue structure over time, business model of airlines such as charter airlines, low-cost carriers (LCC), and full-service carriers (FSC), and geographical aspects are not considered in the present study.

Literature Review

Airline profitability was not as focused when state-owned carriers were supported by government subsidies, and seats and fares were regulated by inter-governmental agreements (The Economist, 2014), which demotivated airlines' voluntary efforts for cost reduction at the industry level. The air transport liberalization movement in 1980s in the U.S. and early 1990s in Europe dramatically changed the financial environment for the airlines and created competition and new business models (Scotti & Volta, 2017). This new wave, coupled with a structural vulnerability to outside shocks (Scotti & Volta, 2015), caused poor financial performance among the airlines (Brugnoli et al., 2015) and airline profitability became one of the major concerns within the industry. As a result, priority emphasis was placed on several different issues related to airline profitability: the relationship between profit fluctuations and industry demand, the cyclical performance of airline earnings, and the relationship between profitability and business models (Scotti & Volta, 2017).

Since the 1980s, published literature on airline profitability has focused primarily on technical efficiency and total factor productivity (Yu, 2016). Studies have traditionally sought to establish the changes in technical efficiency and productivity over time, and to identify which factors have driven such changes (Scotti & Volta, 2017). Other studies have investigated airline productivity and cost competitiveness (Oum & Yu, 1998). However, Heshmati and Kim (2016) note the lack of profitability among airlines is not always due to poor performance alone, and understanding profitability requires an integrative approach with inclusion of all associated factors.

Profitability of the Airline Industry

Despite continuous traffic growth, the airline industry's financial situation has not been regarded as healthy, and demonstrates that increases in traffic volume do not mirror increases in profit (Cronrath, 2017). The world airlines' profit average for decades was almost zero. Precisely, the average net annual profit accumulated from 1978 to 2010 amounts to -$0.04 billion (Cronrath, 2017). Profitability of the airline industry can be expressed with a simple formula stated as (Vasigh, Fleming, & Tacker, 2018):

Profit = OR – OC = (RPM × RRPM) – (ASM × CASM), where:
OR = Operating revenue
OC = Operating costs
RPM = Revenue passenger miles
RRPM = Revenue per revenue passenger mile
ASM = Available seat miles
CASM = Cost per available seat mile

Since every airline's operation and business model are unique, it is difficult to compare operating costs and revenues from airline to airline (Vasigh, Fleming, & Tacker, 2018). In general, airlines' financial structures are largely dependent on the aircraft size, aircraft type, and the length of flights. For instance, while the operating costs of turboprops are much lower than for regional jets, especially on short-haul routes, long-haul routes using wide-body jets are the primary revenue source of major airlines. Therefore, airlines have sought to find an optimal financial structure to increase profits by reducing operating costs and improving the efficiency in the utilization of resources (Belobaba, 2009).

Profitability of the U.S. Airline Industry between 2009 and 2018

In the United States, the airline industry has undergone considerable restructuring for the last decade, primarily due to the recession between 2007 and 2009, the advent of low-cost carriers, and the jet fuel price surge (Zarb, 2018). As a result, profits of the U.S. airline industry have been volatile and forced U.S. airlines to examine profitability of operations. Multiple factors including fuel prices, traffic demand, fare competition, ancillary charges, and exchange rates have been attributed as causes of profit volatility (Martin, 2018; Zarb, 2018). With the rapid growth of low-cost carriers in the U.S. domestic market, the traditional network legacy and regional carriers have experienced intense price competition. Under the severe cost competition and fluctuation of a volatile business environment, the U.S. airline industry has made substantial efforts to optimize cost and revenue structures as well as expand business portfolios for diversifying revenue sources (Claussen, Essling, & Peukert, 2018; Zarb, 2018). This effort has significantly contributed to improved financial performance and differentiated the pricing and cost management strategy of the U.S. airline sector from the previous decades (Luttmann, 2019). According to data from the U.S. BTS (2018a), the 23 largest U.S airlines reported collecting a record of nearly $25 billion in profits in 2015, and have continued profitable operations.

Methodology

Dataset Description

This research project uses secondary data from the U.S. BTS (2018b) Form 41 - Air Carrier Financial Data: Schedule P-1.2, which is published quarterly at publicly accessible United States Department of Transportation website (https://www.transtats.bts.gov/DL_SelectFields.asp). Although the Form 41 requires a uniform system of accounts, each airline employs its own accounting methods and cost allocation schemes. Consequently, the collected cost profiles, or yearly cost trends, may have a difference in cost accounting standards as compared to real differences in operating cost performance (Belobaba, 2009).

Among the different types of airline financial reports found within Form 41, the Schedule P-1.2 reports filed quarterly profit and loss statements during 2009 - 2018 contain 4,136 cases from 94 U.S. airlines which generate annual operating revenues of $20 million or more. The dataset is divided into four groups of airlines: major airlines, national airlines, regional airlines, and all-cargo airlines. Various financial performance indicators are provided, including: operating and non-operating revenues, operating and non-operating expenses, depreciation and amortization, operating profit, income tax, and net income. After careful examination of the dataset of 4,136 cases, the researchers narrowed the scope of the study to include only the major airline group, due to the highly disparate financial structure among the four airline groups. For instance, financial data of all-cargo carriers do not include passenger-related revenues and costs, and regional airlines have too many "zero" values with financial data, which could distort the outcome of the study.

The major airline group consists of 26 large airlines in the U.S. with annual operating revenues of $1 billion or more per airline, representing 1,329 cases in the dataset. While 392 cases indicate a loss, 937 cases indicate generation of a profit. Table 1 shows the major airline group's data profile of the Schedule P-1.2 reports for this study.

Table 1
Schedule P-1.2: Major Airline Group Data Profile between 2009 and 2018

Year	Profit	Loss	Total
2009	85	86	171
2010	99	45	144
2011	83	49	132
2012	73	40	113
2013	102	36	138
2014	105	35	140
2015	114	20	134
2016	106	26	132
2017	106	26	132
2018 (- Sep.)	64	29	93
Total Cases	937	392	1329

Variable Selection

The dataset includes 52 data fields describing the participating airlines' quarterly financial statistics data. Excluding non-operating incomes and expenses,

and aggregate sums for operating revenues and costs, 19 variables were selected for this study as shown in Table 2.

Table 2
Schedule P-1.2: U.S. Airline Operating Profit and Loss Statement Structure, Monetary Value

Variables Name	Measurement Level	Model Role	Mean (,000)	Std. Deviation
Profit/Loss				
Profitability (1: Profit / 0: Loss)	Binary	Target	0.71	0.456
Profit/Loss Value	Interval	Rejected	61715	264329
Operating Revenue				
Scheduled Passenger Revenue (SPR)	Interval	Input	825320	1103041
Mail Revenue (MR)	Interval	Input	2614	4429
Property Revenue-Freight (PRF)	Interval	Input	19513	24705
Property Revenue-Baggage Fee (PRBF)	Interval	Input	26028	44442
Charter Revenue-Passenger (CRPG)	Interval	Input	3971	12690
Charter Revenue-Property (CRPT)	Interval	Input	4.2	24
Reservation Cancellation Fee (RCF)	Interval	Input	19417	30607
Miscellaneous Operating Revenue (MOR)	Interval	Input	29951	75579
Public Service Revenue (PSR)	Interval	Input	320	1791
Transport Related Revenue (TRR)	Interval	Input	193075	454952
Operating Expenses				
Flight Operation Cost (FOC)	Interval	Input	372542	469661
Maintenance Cost (MC)	Interval	Input	97498	126284
Passenger Service (PGS)	Interval	Input	76649	107281
Aircraft and Traffic Service (ATS)	Interval	Input	146923	215600
Promotion and Sales (PRS)	Interval	Input	58997	86624
General Administration Expense (GAE)	Interval	Input	76898	119039
Depreciation and Amortization (DA)	Interval	Input	49978	71563
Transport Related Expenses (TRE)	Interval	Input	142466	344938

Note. For descriptive analysis of the Percentage Value dataset, see Appendix 1.

In this dataset, as shown in Appendix A, examination of the data distribution revealed most of independent variables do not meet a normality assumption. This is due to the seasonal fluctuation of airline financial performance, the gap between 26 airlines' profit or loss scale, disparate financial structure, and the numerous business factors affecting these values. For instance, while American Airlines shows a wide spectrum business portfolio for most of the revenue fields, American Eagle Airlines' business structure is heavily concentrated on passenger transportation revenue. In this case, the Profit/Loss Value variable cannot be used as a target variable because interval values using a linear regression model require meeting the normality assumption. In this regard,

the profitability variable (1: Profit / 0: Loss) is chosen as a target variable because the binary variable runs the logistic regression model, which does not have any required assumptions such as normality or linearity.

In this study, to answer RQ3, the unit of the operating revenue and expenses was reviewed using two different scales of measurement: monetary and percentage values, creating two separate datasets. While the monetary values were used for operating revenue and expense variables in the original dataset, duplicate datasets were created by converting the monetary values into percentage values: dividing each monetary value either by total sum of operating revenue (Monetary Value) or operating expenses respectively and multiplying the decimal by 100 (Percentage Value).

Data Analysis Method and Justification

Due to the large scale of the database and the number of variables, data mining techniques were used for analysis of airline financial data to develop predictive models for estimating the U.S. airlines' operating profit and loss structure. A large number of modeling techniques are labeled "data mining" techniques (Kim, 2008). Predictive modeling is the process of applying a data mining algorithm or statistical model to data in order to predict future observations (Shmueli, 2010). Model performance can be assessed through the use of cumulative lift charts and receiver operating characteristic charts.

Among various data mining techniques, regression analysis and decision tree analysis have many applications for development of a predictive model in a wide spectrum of industries (Halili & Rustemi, 2016). While an artificial neural network method can also be used to construct prediction models, this method was not considered for this study. The artificial neural network method performs better compared to regression and decision tree methods with a larger number of categorical variables than used in this study (Kim, 2008). Logistic regression provides generality, interpretability and robustness, and reliability can be monitored using statistical indicators (Tuffery, 2011). This method of analysis is utilized as the dependent variable of profitability is binary, is not correlated with the independent variables, and the model has no required assumptions for normality or linearity. Therefore, decision tree and logistic regression methods were selected to predict the major U.S. airline group's profitability and identify key influential factors for profitable operations of the major airline group. After analyzing these factors, the best model for predicting the major airline group's future operation profitability was selected.

Data Analysis Process

Data preparation. In this study, SAS Enterprise Miner v. 14.3 (SAS EM) was used to analyze the datasets. Only variables which represent revenue or

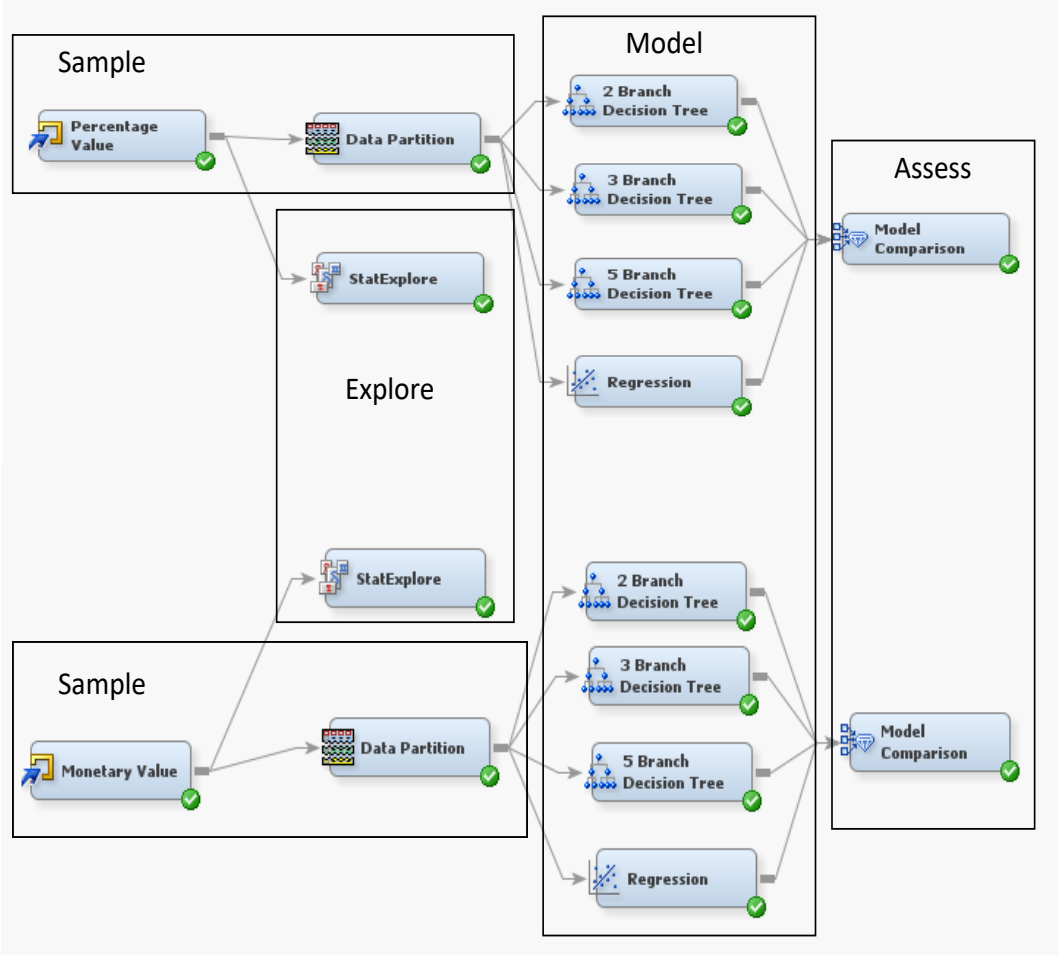
expenses were included in the analyses; Net Income, Carrier, Carrier Group, Carrier Name, Quarter, Year, and Region were excluded. Depending on airlines' operation structure and accounting practices, any specific revenue or cost item could have been intentionally left blank. After careful examination of the individual cases within the dataset, input variables with missing cost values such as operating revenue or expenses were found as non-financial gain/loss so were treated as a "0" value instead of a missing value.

As the first step of the analysis, using SAS EM's Explore function, visual investigation of the dataset was conducted by reviewing individual histograms for each of the interval variables. Information presented in SAS Explore displays the mean values, maximum/minimum for the interval variables, and the number of class levels, missing values, and modal value for class variables. Further examination of distributions of the 18 interval variables found that 10 variables have skewed distributions greater than 3.0, which are listed in Appendix A. As both logistic regression and decision trees methods do not strictly require the assumptions for normality or linearity, these variables are not transformed. Neither dataset displayed missing values.

Data mining: SEMMA process. Data mining is an iterative process where answers to initial questions can lead to more interesting and specific questions (SAS Institute, 1998). To provide an effective methodology for the process operations, SAS Institute presented the SEMMA process which divides data mining into five stages: sample, explore, modify, model and assess. Based on the SEMMA process, the SAS Enterprise Miner software is configured to provide an end-to-end business solution for data mining (SAS Institute, 1998).

This study followed the SEMMA process, as shown in Figure 1. First, in the "sample" stage, the dataset was reviewed and divided into two equal-sized partitions: a training group (50%) and a validation group (50%). In the "explore" stage, each variable was reviewed using the StatExplore function of SAS EM to examine missing values, distribution, and statistical profiles. In the "modify" stage, the dataset was not imputed or transformed as there was no assumption requirement for normality or linearity, and no missing values.

In the "model" stage, three decision trees and logistic regression models were developed to predict the probability that a profit or loss would be observed. In order to compare the various options that separate the individuals of each class, three different trees with 2, 3, and 5 branch options were built using an automatic pruning and average squared error (ASE) method. The three decision tree models were then compared with the outcome of the regression model in the "assess" stage. The ASE was chosen for the decision tree model because of the binary target variable and estimate predictions for this model.



F*igure 1*. The SEMMA process map. A Modify step was not conducted in this study.

Results

Decision Tree Models

Three decision tree models were built for both the Percentage and Monetary Value datasets with 2-, 3-, and 5-branch options using the Decision Tree nodes in SAS EM. The appropriate number of branches were selected in the Splitting Rule option, and the average square error (ASE) was selected as the subtree assessment measure. Outcomes from these models are summarized in Table 3 and the detailed tree maps can be found in Appendix B. The Percentage Value 2-branch model has the lowest ASE value, and is the best model among the six decision tree models.

Table 3
Comparison of the Six Decision Tree Structures and Hierarchy

Dataset	Decision Tree Model	Predictor Factors		
		Level-1	Level-2	Level-3
Percentage Value	2 Branches	Freight Revenue	Misc. Operation Revenue	Public Service Revenue
				General Admin. Expenses
			Transport-related Expenses	Flying Operation Costs
	3 Branches	Freight Revenue	Misc. Operation Revenue	Passenger Service Costs
			Transport-related Expenses	Mail Revenue
				Charter Revenue (Passenger)
				Property Revenue (Baggage)
	5 Branches	Freight Revenue	Promotion and Sales Costs	Transportation Revenue
				Transportation Expenses
			Misc. Operation Revenue	Passenger Service Costs
Monetary Value	2 Branches	Transportation Expenses	Property Revenue (Baggage)	Public Service Revenue
				Transportation Expenses
	3 Branches	Transportation Revenue	Transport-related Expenses	Depreciation & Amortization
				Transportation Revenue
			Freight Revenue	
	5 Branches	Transportation Revenue	Transport-related Expenses	Depreciation & Amortization
				Promotion and Sales Costs
			Freight Revenue	

The tree map and windows display two different views of the decision tree (SAS Institute, 2013). While the tree window presents a standard decision tree, the tree map window uses the width bands to illustrate the proportion of observations in each node in the row. The root node of the 2-branch decision tree for the Percentage Value dataset indicates 70.5% of the cases in the dataset for major airlines reported a profit, while 29.5% reported a loss. Under the top level, the factors are listed in order of importance. These factors contributed to the prediction of profit or loss in the following order:

1. The most important variable is PRF, it further differentiates between cases which reported profits and those which reported losses. Among cases that

have had PRF less than 5.4%, 74.9% will turn a profit: Among cases equal or greater than 5.4%, only 40.7% will turn a profit.

2. The second most important variable is MOR. Among cases which showed equal or greater than 5.7% in MOR, 92.7% showed a profit, while among cases with less than 5.7% in MOR or had missing values, only 72.6% showed a profit.
3. Among cases with equal or greater than 5.7% in MOR, GAE was the next most important variable. Among cases with less than 21% in these expenses, 93.9% earned profits. PSR was the most influential variable among cases which showed less than 5.7% in MOR. For cases with equal or greater than 0.16% revenue from public services, 94.9% showed a profit.

Among the Monetary Value decision tree models, the 3-branch decision tree has the lowest ASE value. As shown in Appendix B, the root node indicates 70.4% of the cases in the Monetary Value dataset for major airlines reported a profit, while 29.5% reported a loss. Under the top level, the factors are listed in order of importance. These factors contributed to the prediction of profit or loss in the following order:

1. The most important variable is TRR, which further differentiates between cases which reported profits and those which reported losses. Among cases that showed TRR less than $424,926,500, 69.8% will earn a profit: Among cases with TRR between $424,926,500 and $622,897,700, only 33.3% of those cases earned a profit, while among airlines with TRR equal to or greater than $622,897,500, 85% earned profits.
2. The second most important variable among the cases with TRR between $424,926,500 and $622,897,700 is PRF. Cases with less than $32,298,500 or missing values for PRF showed only 20% earning a profit, while among cases with $32,296,600 or greater values of PRF, 100% earned a profit.
3. The second most important variable among the cases with TRR less than $424,926,500 is TRE. 71.5% of cases with less than $84,706,000 or with missing values for TRE were observed to earn a profit, while only 37.9% of cases with TRE between $84,706,000 and $115,542,000 showed a profit. For cases with TRE greater or equal to $$115,542,000, 70.8% earned a profit.

The six decision tree models show different tree structures and hierarchy of decision factors, as compared in Table 3. While the Freight Revenue variable was chosen in the Percentage Value dataset as the most important predictor factor for airline profitability, the decision tree models with the Monetary Value dataset indicated Transportation Related Revenue and Expenses as the most influential to predict profitability.

The Leaf Statistics window in SAS EM presents the classification rates of training and validation samples for each leaf node in the tree model. The leaf statistics indicate the frequency percentages for each target class level within each leaf from the training and validation data. Leaf statistics charts for selected models' decision trees are found in Appendix C.

Cumulative lift charts indicate the percentile on the x-axis and the lift on the y-axis, where the default (no model) is a horizontal line intersecting the y-axis at 1. The higher the lift index, the better the model. Cumulative lift charts for selected models in this study are shown in Appendix D. The Percentage Value 2-branch decision tree cumulative lift chart appears to indicate the highest lift index of the models developed in this study, as shown in Figure 2. Using this chart, predicted probability of profit can be calculated by multiplying the lift value and depth as percentage value. The chart in Figure 2 indicates that at 20% of depth, the lift value is approximately 1.3, which means that if a random 20% of the cases are observed, there is a 26% chance the case will show a profit. Generally, lift value decreases as the selected proportion of the data increases (SAS Institute, 2013) and holds true for this model.

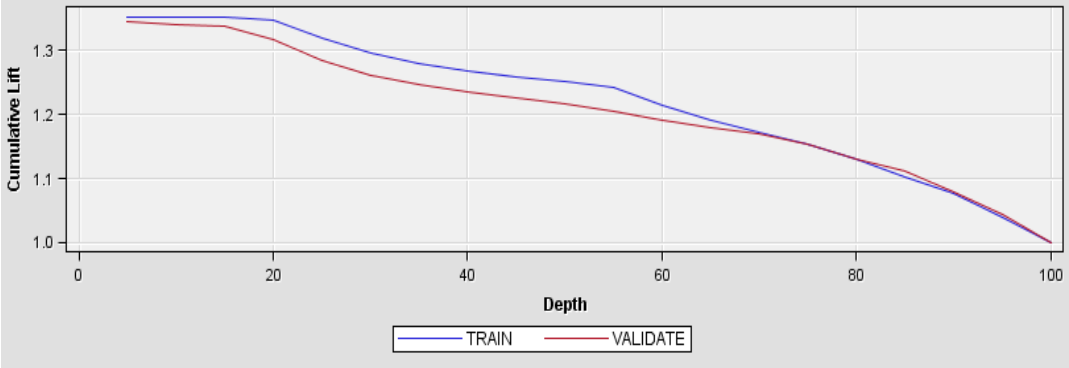


Figure 2. Percentage value 2-branch decision tree cumulative lift chart.

Subtree assessment plots for all decision tree models were observed, and are found in Appendix E. These plots show the ASE of corresponding to each tree leaf in the sequence. The Percentage Value dataset showed trees with an optimal range between 11 to 19 leaves for the lowest ASE when viewing the validate results. The Monetary Value dataset showed optimal ranges of between 7 and 20 leaves generating the lowest ASE values. All of the plots indicated the performance of the training sample becomes better as the tree becomes more complex. The performance of the validation sample only improves to a certain number of leaves and then performance decreases as the complexity of the model increases.

Variable importance for the Percentage Value dataset is found in the Output from SAS EM, and is displayed in Table 4. Variable importance is an indication of which predictors are most useful for predicting whether an airline will experience a profit or loss, the target variable. Results indicate the most important variables for the Percentage Value dataset decision trees are PRF, MOR, and PRS, with Property Freight appearing among the first four variables in all models.

Table 4
Variable Importance in 2-, 3-, and 5-Branch Decision Trees for Percentage Value Dataset

Dataset Model Variable Name	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
2-Branch Decision Tree				
Prop_Freight	1	1.0000	1.0000	1.0000
Trans_Expenses	2	0.9909	0.5922	0.5977
Prop_Bag	1	0.7349	0.9856	1.3410
Misc_Op_Rev	1	0.6825	0.5174	0.7581
Promotion_Sales	1	0.6736	0.2111	0.3134
Pub_Svc_Revenue	1	0.6363	0.4823	0.7581
Trans_Revenue	1	0.6242	0.4064	0.6511
Flying_Ops	1	0.5603	0.6621	1.1817
General_Admin	1	0.4496	0.1994	0.4434
3-Branch Decision Tree				
Misc_Op_Rev	2	1.0000	0.3549	0.3549
Prop_Freight	1	0.9329	1.0000	1.0719
Trans_Revenue	1	0.8692	0.7890	0.9078
Prop_Bag	1	0.8102	0.0000	0.0000
Trans_Expenses	1	0.8011	0.6448	0.8050
General_Admin	1	0.6530	0.4634	0.7097
Pax_Service	1	0.6243	0.2647	0.4240
Maintenance	1	0.5920	0.4464	0.7540
Charter_Pax	1	0.5240	0.0000	0.0000
Mail	1	0.4280	0.1198	0.2800
5-Branch Decision Tree				
Promotion_Sales	1	1.0000	0.0000	0.0000
Prop_Freight	1	0.8257	1.0000	1.2111
Trans_Expenses	2	0.7812	0.8114	1.0387
Misc_Op_Revenue	1	0.7129	0.3549	0.4979
Pax_Service	1	0.5526	0.2647	0.4790
Trans_Revenue	1	0.4720	0.4651	0.9853

Variable importance for the Monetary Value dataset is found in the Output from SAS EM, and is displayed in Table 5. Results indicate the three most

important variables for the Monetary Value dataset decision trees are Transport Related Expenses (TRE), Transportation Related Revenue (TRR), and PRBF. These variables are found in the top three influencing variables for models of 2-, 3-, and 5-branches.

Table 5
Variable Importance in 2-, 3-, and 5-Branch Decision Trees for Monetary Value Dataset

Dataset Model Variable Name	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
2-Branch Decision Tree				
Trans_Expenses	2	1.0000	0.7414	0.7414
Prop_Bag	1	0.9885	1.0000	1.0117
Trans_Rev_Pax	1	0.8886	0.5690	0.6403
Charter_Pax	1	0.7509	0.2794	0.3721
Pub_Svc_Revenue	1	0.7247	0.3350	0.4623
3-Branch Decision Tree				
Trans_Revenue	2	1.0000	0.7923	0.7923
Trans_Expenses	2	0.8387	1.0000	1.1923
Prop_Bag	2	0.7758	0.7428	0.9574
Maintenance	1	0.5825	0.6652	1.1420
Deprec_Amort	1	0.5738	0.0000	0.0000
Promotion_Sales	1	0.4961	0.6422	1.2944
Trans_Rev_Pax	1	0.4608	0.1956	0.4246
Flying_Ops	1	0.3916	0.2399	0.6126
Prop_Freight	1	0.3058	0.3149	1.0300
5-Branch Decision Tree				
Trans_Expenses	2	1.0000	1.0000	1.0000
Trans_Revenue	1	0.9230	0.5449	0.5903
Prop_Bag	1	0.7975	0.9287	1.1645
Promotion_Sales	2	0.6891	0.6065	0.8802
Deprec_Amort	1	0.6306	0.0000	0.0000
Flying_Ops	1	0.4897	0.4595	0.9382
Prop_Freight	1	0.3372	0.2974	0.8820

Logistic Regression Model

Logistic regression models were used to evaluate the Percentage and Monetary Value datasets using the Regression node in SAS EM. The entry and stay significance levels were changed from the default value of 0.05 to 0.1 to allow additional variables to be included in the model, and the maximum number of steps was changed to 20.

The Fit Statistics results for the logistic regression models are found in Table 6. Although the similarity in values between training and validation rates may indicate consistency and validity of results, the ASE values for the logistic regression models are higher than the values for the decision trees. The ASE values for the logistic regression model for the Value dataset are the highest among all ASE values. The lower average square error means the model performs better as a predictor because it is "wrong" less often (SAS, 2010). It appears neither regression model is an obvious better predictor.

Table 6
ASE Values for the Target Variable Profit_Loss from SAS EM Fit Statitistics Output:

Model	Data Group	
	Train	Validation
Percentage Value		
Decision Tree – 2 Branches	0.157112	0.161899
Decision Tree – 3 Branches	0.139775	0.179158
Decision Tree – 5 Branches	0.155727	0.182018
Logistic Regression	0.18565	0.175882
Monetary Value		
Decision Tree – 2 Branches	0.176161	0.178845
Decision Tree – 3 Branches	0.124499	0.171524
Decision Tree – 5 Branches	0.138679	0.174849
Logistic Rgression	0.202155	0.200348

The cumulative lift chart for logistic regression also indicates the percentile on the x-axis and the lift on the y-axis, where the default (no model is a horizontal line intersecting the y-axis at 1. The cumulative lift charts for the linear regression models are found in Appendix D, and are interpreted similarly to the cumulative lift charts for the decision tree models.

The Analysis of Maximum Likelihood Estimates table in the Output indicates the variables included in the final model selected by SAS EM. These variables are displayed in Table 7. The estimate, or correlation coefficient shows the sign of the effect of each predictor, and if it has a positive or a negative effect on the outcome. In the Percentage Value model, both PRF and TRE variables have negative effects on the profitability of an airline. In the Monetary Value model, both Charter Revenue Property revenue and MOR have positive effects on the profitability of an airline.

The exponential value indicates magnitude of the effect. For example, the exponential value for the PSR in the Percentage Value dataset has the largest effect with an exponential value of 2.781. Thus, if the value of this variable is increased by two units, the probability that a case will indicate a profit increases by 78%.

Iteration plots, as found in Appendix F, display the value of a model assessment measure on the vertical axis for different steps in the stepwise process. The plots indicate the optimal number of iterations for the logistic regression models, and confirm findings in Table 7, which suggest 8 steps are the optimal number for the Percentage Value model and 2 steps are optimal for the Monetary Value model.

Table 7
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Std. Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp (Est)
Percentage Value	1						
Intercept	1	-2.0836	0.8396	6.16	0.0131		0.124
Deprec_Amort	1	0.1098	0.0447	6.03	0.0141	0.1443	1.116
Flying_Ops	1	0.0342	0.0134	6.49	0.0109	0.1774	1.035
Misc_Op_Revenue	1	0.1184	0.0323	13.41	0.0003	0.3034	1.126
Pax_Service	1	0.1228	0.0470	6.82	0.0090	0.1710	1.131
Prop_Freight	1	-0.2360	0.0460	26.35	<.0001	-0.3049	0.790
Pub_Svc_Revenue	1	1.0227	0.6380	2.57	0.1089	0.1937	2.781
Trans_Expenses	1	-0.0494	0.0209	5.58	0.0182	-0.2669	0.952
Trans_Revenue	1	0.0645	0.0187	11.94	0.0006	3.895	1.067
Monetary Value	1						
Intercept	1	0.6896	0.0933	54.63	<.0001		1.993
Charter_Prop	1	0.0395	0.0297	1.77	0.1832	0.5591	1.040
Misc_Op_Revenue	1	4.42E-6	1.516E-6	8.51	0.0035	0.2062	1.000

Model Comparison

The Model Comparison node in SAS EM was used to compare the three decision tree models and the logistic regression model for each of the datasets. ASE was chosen as the Selection Statistic, and comparison results are based on validation data. The Fit Statistics window displays several computed statistics for

all partitions of the training and validation data in the decision tree models (SAS Institute, 2013). The ASE values for training and validation samples for each of the each of the models are found in Table 6. The lowest validation values for ASE were observed in the 2-branch decision tree model for the Percentage Value dataset and the 3-branch decision tree model for the Monetary Value dataset, while the lowest training values were observed in the 3-branch decision trees for both models. The discrepancy in values between the training and the validation samples calls into question the consistency and validity of these models, and as such neither were determined to be indicative of an ideal model. The cumulative lift charts and receiver operating characteristic charts and indices were examined to identify a better assessment method to select the best prediction model.

The comparative cumulative lift charts for the decision trees and the logistic regression models are found in Appendix G. For both datasets, the 3-branch decision tree models appear to be the superior models, as they show the higher lift indexes overall. The indications from the cumulative lift charts are inconclusive when compared other tests.

The receiver operating characteristic (ROC) chart displays values of the true positive fraction on the vertical axis, and the false positive fraction on the horizontal axis; it displays the tradeoff between sensitivity and specificity. The straight diagonal line is the baseline; the larger the area between the ROC chart of the model and the diagonal line, the better the model. A perfect test has an area or ROC index of 1, thus the higher the value, the more accurate the model (Statistics How To, 2019).

As seen in Figure 3, the ROC chart visually demonstrates the larger area under the logistic regression line (brown line) until approximately 0.3 on the x-axis (specificity), after which the 2-branch decision tree line (the green line) demonstrates the larger area. The full ROC charts for training and validation samples for both datasets are found in Appendix H, the ROC indices are shown in Table 8. Based on ROC indices, the best model appears to be the 3-branch decision tree using the Monetary Value dataset, followed by the 3-branch decision tree using the Percentage Value dataset, and then the 5-branch decision tree using the Monetary Value dataset.

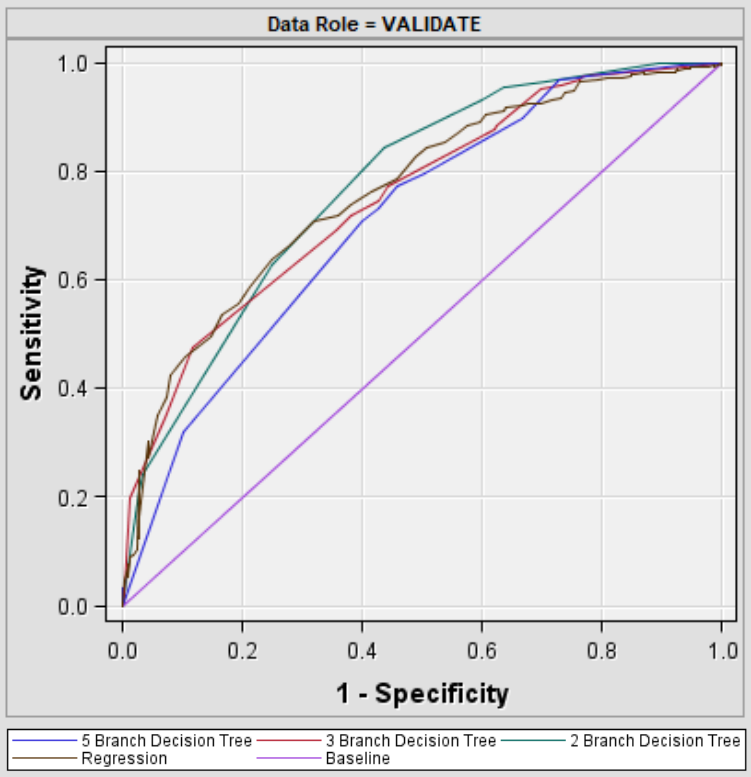


Figure 3. ROC chart comparing decision tree models and logistic regression model for percentage value dataset.

Table 8
ROC Index values from SAS EM Model Comparison Node Output:

Ananlsys Models	ROC Index
Percentage Value	
Decision Tree – 2 Branches	0.79
Decision Tree – 3 Branches	0.84
Decision Tree – 5 Branches	0.79
Logistic Regression	0.76
Monetary Value	
Decision Tree – 2 Branches	0.68
Decision Tree – 3 Branches	0.85
Decision Tree – 5 Branches	0.82
Logistic Regression	0.55

The model selected by the SAS EM Model Comparison node for the Percentage Value dataset is the 2-branch decision tree, then the logistic regression, then the 3-branch decision tree, then the 5-branch model. Among the Monetary Value dataset, SAS EM selected the best model as the 3-branch decision tree, then the 5-branch decision tree, then the 2-branch decision tree, then the logistic regression model. These models are ranked by ASE based on the results of the validation data.

In summary, comparing the eight presented models using two interrelated datasets, it appears the decision tree models are better predictors of profitability for major airlines. Based on ASE values, the 2-branch decision tree using the Percentage Value dataset was identified as the best prediction model, followed by the 3-branch Monetary Value dataset decision tree model. In the 2-branch Percentage Value decision model, PRF is indicated to be the most influential factor, followed by TRE and PRBF. For the 3-branch Monetary Value dataset decision tree model, TRR is indicated to be the most influential factor to predict an airline's profitability, followed by TRE, and then PRBF.

Discussion, Conclusions, and Recommendations

Three decision tree models and one logistic regression model were developed for two datasets to predict potential profit and loss for major airlines using data from the BTS. Decision trees are popular tools due to their relative power, ease of use, robustness with a variety of data and levels of measurement, and ease of interpretability (deVille & Neville, 2013), however decision trees have several limitations. The primary disadvantage is that they can be subject to overfitting and underfitting, especially when using a small dataset, which can limit the generalizability and robustness of the models (Song & Lu, 2015).

This study answered the following research questions:

RQ1: What is the most influential factor among airline revenue and expense in predicting an airline's profitability, using monetary and percentage values as two scales of measure of input variables?

The TRR and TRE were found to be the two most influential factors in predicting the U. S. airlines' profitability. The TRR and TRE are incidentals to the air transportation services performed by airlines (BTS, 2018b). Examples are ancillary passenger services such as Wi-Fi, duty-free, and food and beverage sales and, revenues and expenses from associated ground businesses like ramp operations, aircraft maintenance, refueling, catering, etc. Today, under extreme cost competition and increasing operation costs, the growth of ancillary revenue has positive ramifications that significantly benefit airlines' financial performance and yielding operation margin (Warnock-Smith, O'Connell, & Maleki, 2017). In this regard, the result of this study highlights and supports the importance of the

Transportation Related Revenue and Expenses for profitability of the U.S. major airlines.

RQ2: Between the decision tree and regression models, which provides the best model for predicting the profitability of a major U.S. airline, using monetary and percentage values as two scales of measure of input variables?

Generally, the decision trees yielded better models than the regression models although no one model indicated obvious superior predictive ability based on the assessments. Based on ASE values, the best predictive model was the 2-branch Percentage Value decision tree, followed by the 3-branch decision tree using the Monetary Value dataset, followed by the 5-barnch decision tree using the Monetary Value dataset, and then the logistic regression model using the Percentage Value dataset.

RQ3: Between a monetary and percentage scale of input variables, which scale can better predict airline profitability?

Neither the monetary or the percentage scale yielded a clear and consistent champion model for predictive modeling. Depending on the assessment method utilized, either the monetary or percentage scale resulted in a better model. The decision tree approach should be more fully explored to exhaust all possibilities of finding a better predictor model. Adjusting the pruning settings to find the number of leaves which provide the best result, changing the assessment measures, or other adjustments may be made to find a better predictive model. Exploring other assessment options may also provide evidence of a better predictive model.

A final assessment of the model should be conducted on a separate dataset, such as BTS data from different years in order to determine the appropriateness of fit. Similar models could be tested by airlines to determine additional variables to be included in the analysis, and the most appropriate model could be used by airlines to predict profit or loss under conditions similar to those tested in the algorithms. In the present study, geographical location and airline business model were not considered as influential predictive factors in airlines' profitability. Indeed, various distinctive business practices of FSC and LCC can play a key role to separate the profitability prediction model between two airline groups due to the disparate cost and revenue structure (Azadian & Vasigh, 2019). Repeat studies can be used to refine the models to provide better predictive analysis for profitability and the factors influencing it. With further improvement, models could potentially be used to predict profit or loss performance based on identified influencing variables.

References

An, B., Chen, H., Park, N., & Subrahmanian, V. (2017). Data-driven frequency-based airline profit maximization. *ACM Transactions on Intelligent Systems and Technology (TIST), 8*(4), 1-28. doi:10.1145/3041217

Azadian, F., & Vasigh, B. (2019). The blaring lines between full-service network carriers and low-cost carriers: A financial perspective on business model convergence. *Transport Policy, 75*, 19-26. doi:10.1016/j.tranpol.2018.12.012

Belobaba, P. (2009). *Airline operating costs and measures of productivity*. (pp. 113-151). Chichester, UK: John Wiley & Sons, Ltd. doi:10.1002/9780470744734.ch5

Brugnoli, A., Button, K., Martini, G., & Scotti, D. (2015). Economic factors affecting the registration of lower CO2 emitting aircraft in Europe. *Transportation Research. Part D: Transportation Environment, 38*, 117-124

Claussen, J., Essling, C., & Peukert, C. (2018). Demand variation, strategic flexibility and market entry: Evidence from the U.S. airline industry. *Strategic Management Journal, 39*(11), 2877-2898. doi:10.1002/smj.2940

Cronrath, E. (2017). *The airline profit cycle: A system analysis of airline industry dynamics*. London: Taylor and Francis. doi:10.4324/9781315188720

International Air Transport Association (2017, December). *Strong airline profitability continues in 2018.* (Press Release No. 70). Retrieved from https://www.iata.org/pressroom/pr/Pages/2017-12-05-01.aspx

Kim, Y. S. (2008). Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. *Expert Systems with Applications, 34*(2), 1227-1234. doi:10.1016/j.eswa.2006.12.017

Luttmann, A. (2019). Evidence of directional price discrimination in the U.S. airline industry. *International Journal of Industrial Organization, 62*, 291-329. doi:10.1016/j.ijindorg.2018.03.013

Martin, H. (2018). *Demand for air travel is strong, but higher fuel costs are cutting into airline profits*. Los Angeles, CA: Tribune Interactive, LLC.

Miranda, U. (2015). *The relationship between terrorism, oil prices, and airline profitability* (Doctoral dissertation). Available from Walden Dissertations and Doctoral Studies Collection at ScholarWorks.

Oum, T. H., & Yu, C. (1998). *Winning airlines: Productivity and cost competitiveness of the world's major airlines.* Boston: Kluwer Academic Publishers.

SAS Institute Inc. [SAS]. (2013). *Data mining and SEMMA.* Data Mining Using SAS® Enterprise Miner™: A Case Study Approach. (3rd Ed.). Retrieved from

https://support.sas.com/documentation/cdl/en/emcs/66392/PDF/default/emcs.pdf

Scotti, D., & Volta, N. (2015). An empirical assessment of the CO2-sensitive productivity of European airlines from 2000 to 2010. *Transportation Research Part D, 37*, 137-149. doi:10.1016/j.trd.2015.04.009

Scotti, D., & Volta, N. (2017). *Profitability change in the global airline industry. Transportation Research Part E, 102*, 1-12. doi:10.1016/j.tre.2017.03.009

Shmueli, G. (2010). To explain or predict? *Statistical Science, 25*(3) 289-310. doi:10.1214/10.STS330

The Economist. (2014). *Why Airlines make such meagre profits*. Retrieved from http://www.economist.com/blogs/economist-explains/2014/02/economist-explains-5

Tuffery, S. (2011). *Data mining and statistics for decision making*. Chichester, UK: John Wiley & Sons, Ltd.

Uniform System of Accounts for Large Certificated Air Carriers, Section 22 General Reporting Instructions, 14 CFR § 241.22 (1972).

U.S. Bureau of Transportation Statistics (2018a). 2016 *Annual and 4th quarter airline financial data*. (Press Release No. BTS 23-17). Retrieved from https://www.bts.gov/newsroom/2016-annual-and-4th-quarter-airline-financial-data

U. S. Bureau of Transportation Statistics (2018b). *Air carrier financial: Schedule P-6.* Retrieved from https://www.transtats.bts.gov/Fields.asp

Vasigh, B., Fleming, K., & Tacker, T. (2018). *Introduction to air transport economics: From theory to applications* (3rd ed.). London: Routledge, Taylor & Francis Group.

Warnock-Smith, D., O'Connell, J. F., & Maleki, M. (2017). An analysis of ongoing trends in airline ancillary revenues. *Journal of Air Transport Management, 64*, 42-54. doi:10.1016/j.jairtraman.2017.06.023

Yu, C. (2016). *Airline productivity and efficiency: Concept, measurement, and applications*. (pp. 11-53) Emerald Group Publishing Limited. doi:10.1108/S2212-160920160000005002

Zarb, B. J. (2018). Liquidity, solvency, and financial health: Do they have an impact on U.S. airline companies' profit volatility? *International Journal of Business, Accounting and Finance (IJBAF), 12*(1), 42.

Appendix A

Interval Variable Summary Statistics from SAS Explore (Train)*

Variable	Role	Mean	Standard Deviation	Min.	Median	Max.	Skewness	Kurtosis
Percentage Value								
Aircrft_Services	Input	13.8129	4.256863	-0.38358	13.91871	33.32241	0.102985	1.764876
Charter_Pax	Input	0.88754	6.223123	-0.06557	0	78.52746	10.12257	106.9473
Charter_Prop	Input	0.000756	0.003232	0	0	0.020446	4.232029	16.74639
Deprec_Amort	Input	4.985672	2.31975	-1.861	5.083964	18.30271	0.75226	2.612824
Flying_Ops	Input	41.73493	9.1982	19.39355	40.95688	74.89001	0.317901	-0.0576
General_Admin	Input	7.893132	4.118253	0	7.089036	54.74681	3.050945	25.42091
Mail	Input	0.206036	0.328721	-0.15974	0.011466	1.822397	1.918964	3.611999
Maintenance	Input	12.23226	7.632439	1.825764	9.916573	44.15967	2.108347	4.081153
Misc_Op_Rev	Input	2.56464	4.51298	-1.04736	0.777581	27.38849	3.193917	11.2518
Pax_Service	Input	7.328825	2.537886	0.506457	7.091891	15.09019	0.402961	-0.38394
Promotion_Sales	Input	5.325785	2.785099	-0.01765	5.878913	14.6952	-0.61206	0.081797
Prop_Bag	Input	2.783436	3.732256	-0.00444	2.130789	20.85931	2.681544	8.076867
Prop_Freight	Input	1.938613	2.65544	-0.48809	0.826041	33.48414	2.635332	17.11991
Pub_Svc_Revenue	Input	0.059317	0.327104	0	0	3.716797	7.070536	53.88385
Res_Cancel_Fees	Input	53.88385	1.514841	-0.13539	1.90206	5.617328	-0.26969	-0.80805
Trans_Expenses	Input	6.686196	9.528534	-0.04055	0.778207	37.12724	1.216252	0.042366
Trans_Revenue	Input	9.207313	10.62012	-1.1096	4.330472	57.90231	1.162911	0.327432
Trans_Rev_Pax	Input	80.83751	13.97299	16.47104	83.92677	100.0457	-0.83092	0.658153
Monetary Value								
Aircrft_Services	Input	146923.7	215599.6	-27	63696	1102324	2.300803	4.80847
Charter_Pax	Input	3970.792	12689.52	-395	0	130539	4.962053	29.10975
Charter_Prop	Input	4.204665	23.94475	0	0	186	6.060885	35.97252
Deprec_Amort	Input	49978.29	71563.15	-463	24008	438220	2.462265	6.301083
Flying_Ops	Input	372542.1	469660.6	51	199764	2819894	1.961089	3.362539
General_Admin	Input	76898.25	119039.3	0	31718	841454	3.059821	11.8463
Mail	Input	2614.391	4428.565	-1033	26	28986	2.210275	5.778444
Maintenance	Input	97497.56	126284	37	56778	624613	2.058989	3.672014
Misc_Op_Rev	Input	29950.97	75578.53	-9628	3578	518782	3.916551	16.33922
Pax_Service	Input	76649.29	107281.1	8	31020	631913	2.330208	5.876528
Promotion_Sales	Input	58996.83	86624.42	-52	26380	516884	2.489426	6.857585
Prop_Bag	Input	26027.51	44442.14	-12	7923	259421	2.791317	8.548951
Prop_Freight	Input	19512.9	24704.66	-1479	7167	146761	1.251317	0.919248
Pub_Svc_Revenue	Input	320.3394	1791.245	0	0	25390	8.032842	76.83812
Res_Cancel_Fees	Input	19417.02	30607.34	-110	6851	155378	2.421174	5.916754
Trans_Expenses	Input	142466.5	344938.7	-49	3694	1879880	3.129448	9.364083
Trans_Revenue	Input	193075.2	454952.1	-4625	23487	2291920	3.144126	9.154019
Trans_Rev_Pax	Input	825321	1103041	175	415399	5328614	2.156862	4.214202

* Non-missing values = 1329, Missing values = 0, Monetary value = US$.000

Appendix B

Decision Tree Outputs From Percentage Value and Monetary Value Datasets

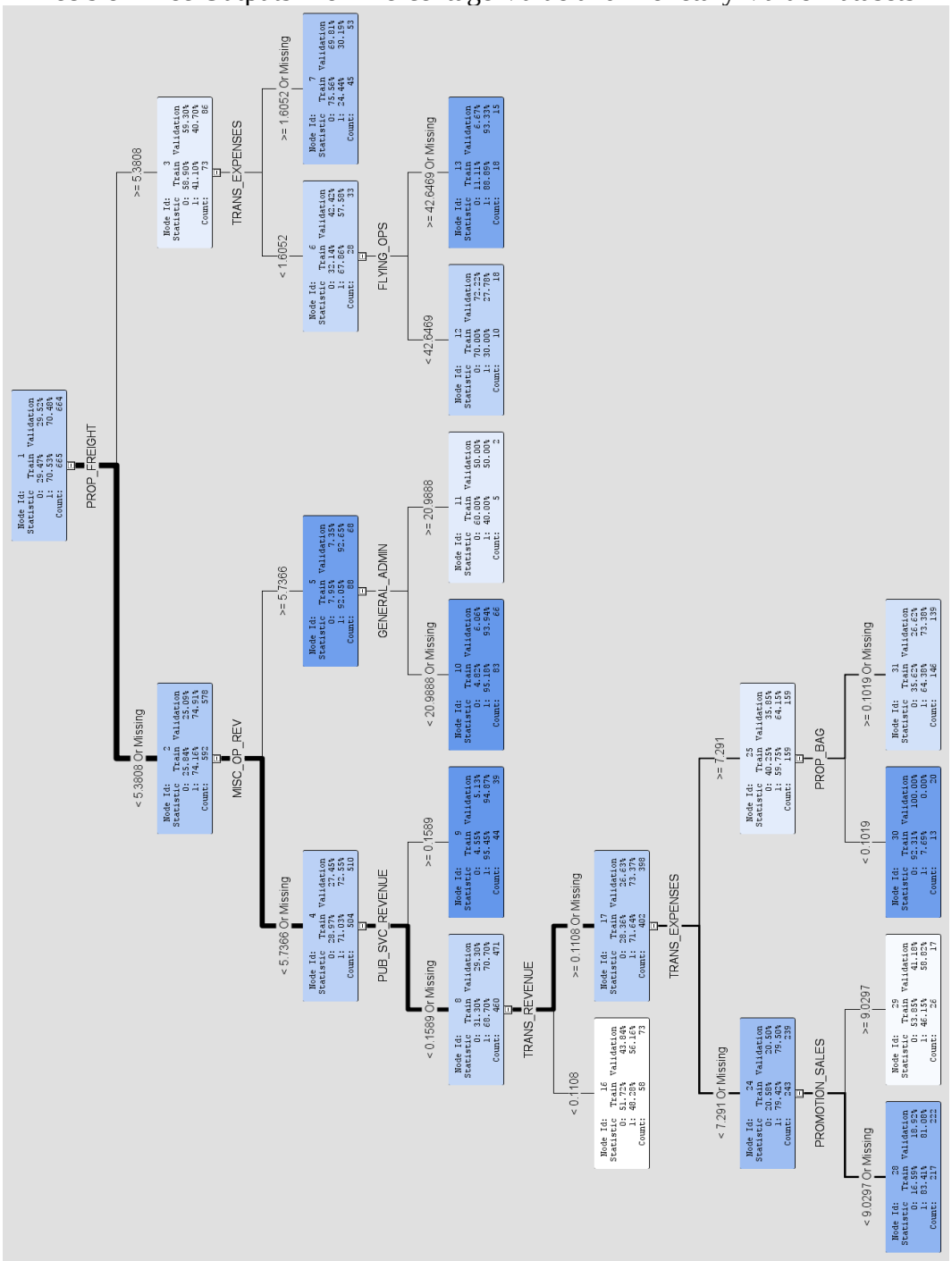


Figure B1. Percentage Value dataset 2-branch decision tree.

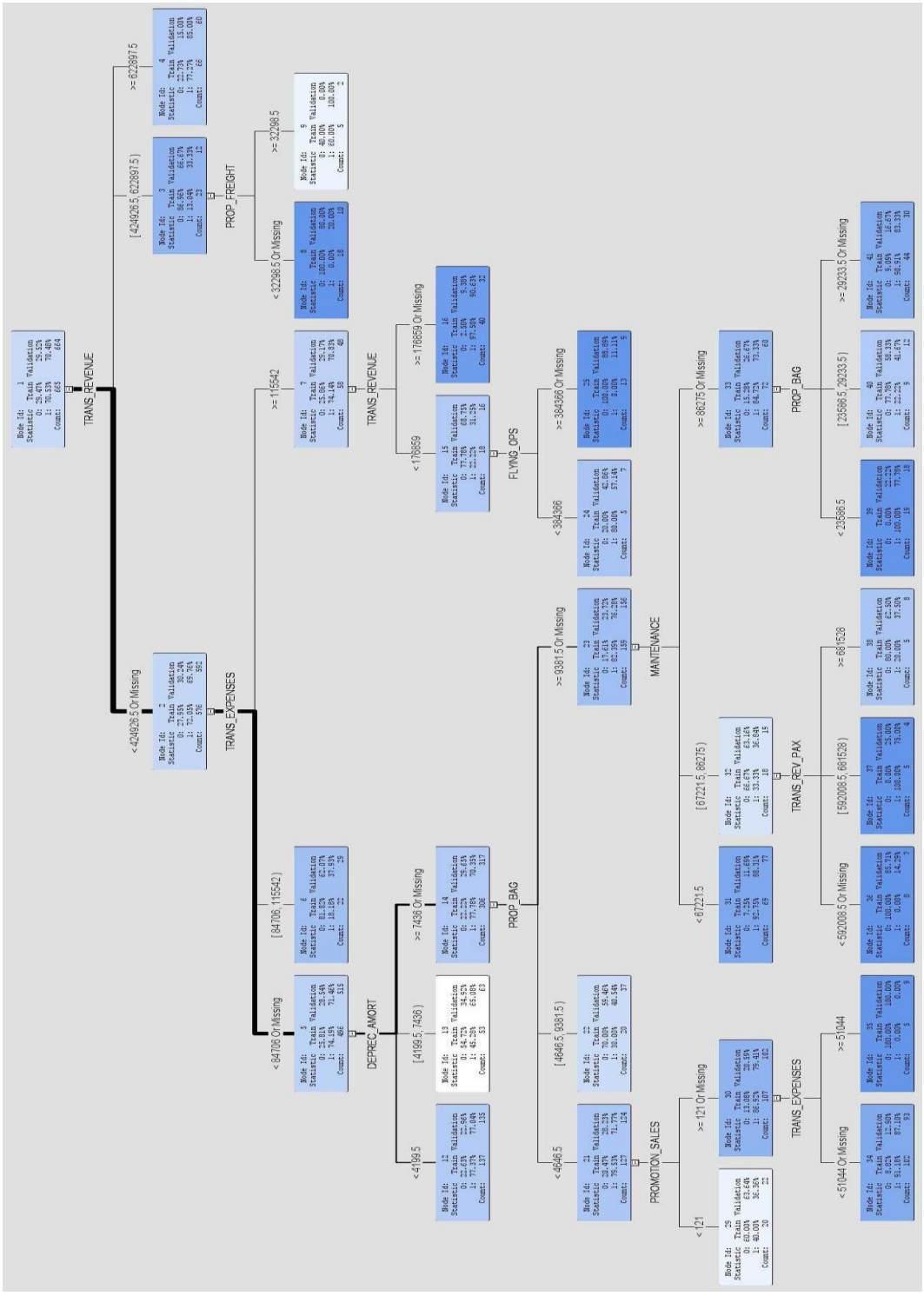


Figure B2. Monetary Value dataset 3-branch decision tree.

Appendix C

Leaf Statistics Charts for Percentage Value and Monetary Value Datasets

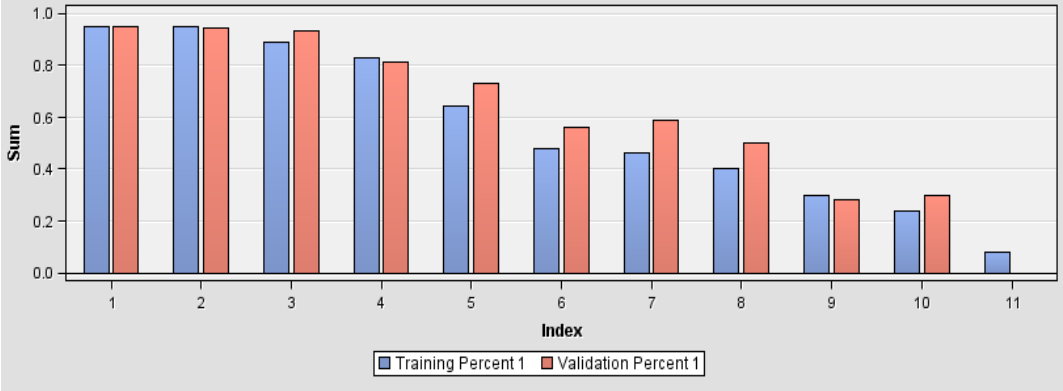


Figure C1. Percentage Value dataset 2-branch decision tree leaf statistics chart.

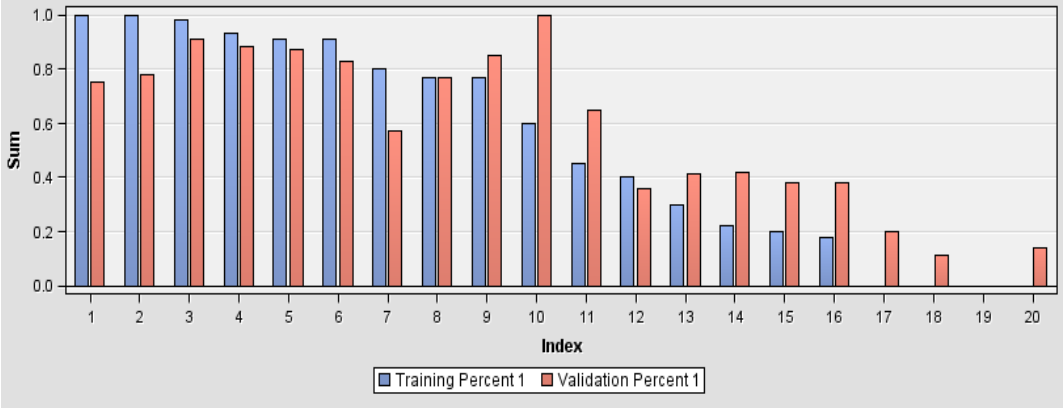


Figure C2. Monetary Value dataset 3-branch decision tree leaf statistics chart.

Appendix D

Cumulative Lift Charts for Percentage Value and Monetary Value Datasets

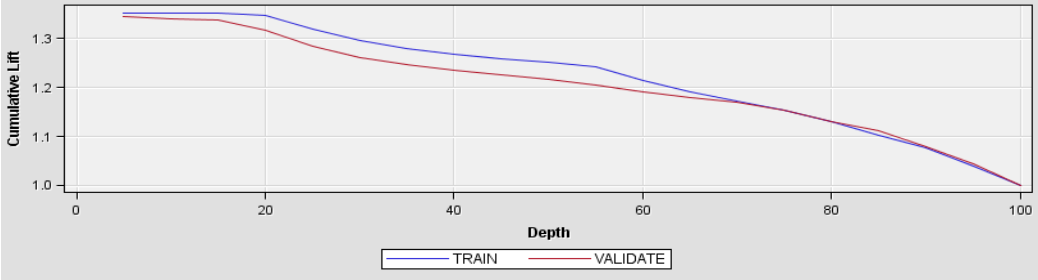


Figure D1. Percentage Value dataset 2-branch decision tree cumulative lift chart.

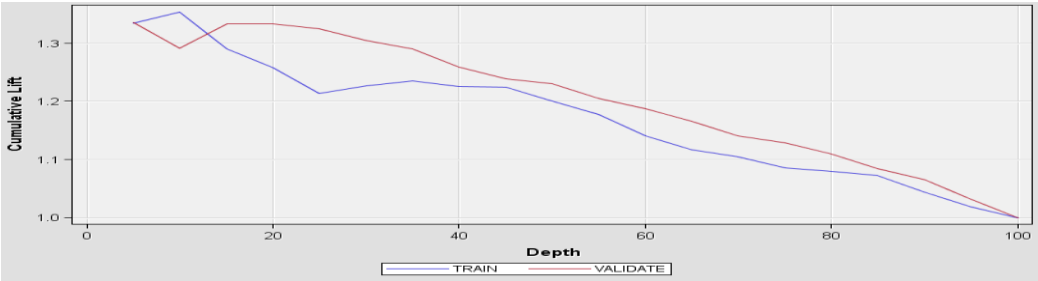


Figure D2. Percentage Value dataset regression model cumulative lift chart.

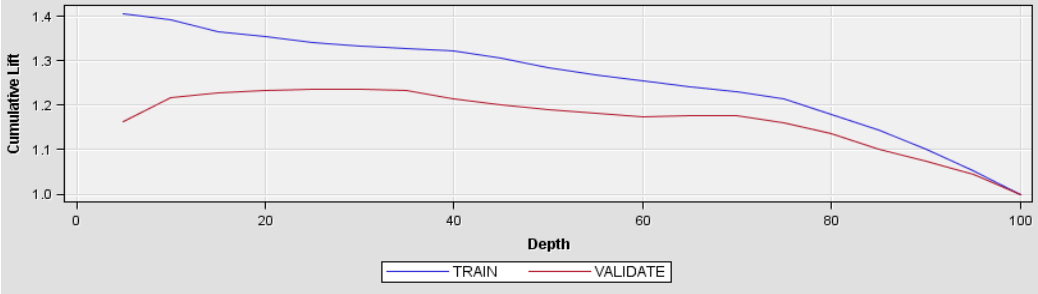


Figure D3. Monetary Value dataset 3-branch decision tree cumulative lift chart.

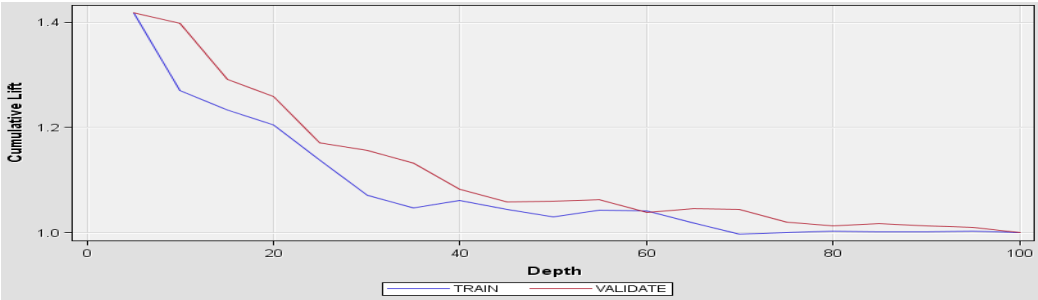


Figure D4. Monetary Value dataset regression model cumulative lift chart.

Appendix E

Subtree Assessment Plots for Percentage and Monetary Value Dataset

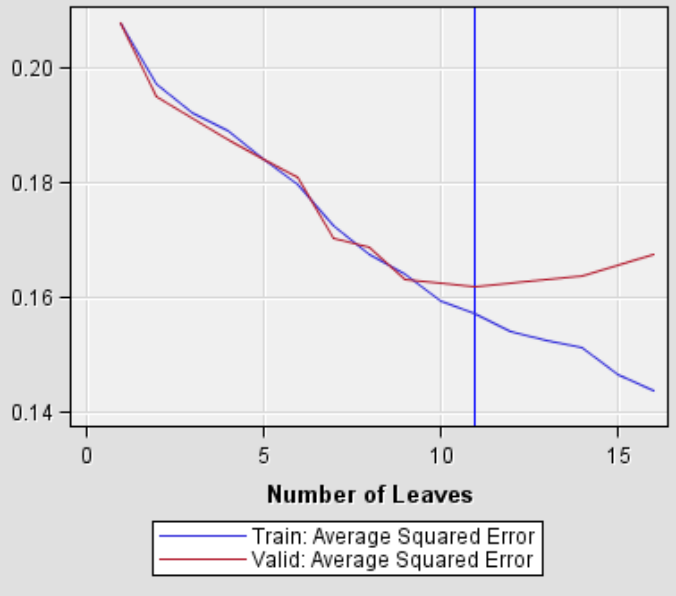


Figure E1. Percentage Value 2-branch decsion tree subtree assessment plot.

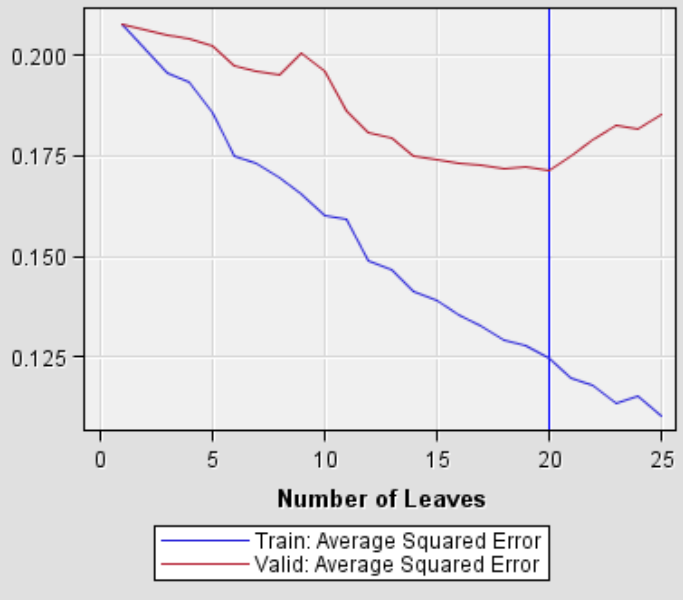


Figure E2. Monetary Value 3-branch decision tree subtree assessment plot.

Appendix F

Iteration Plots for Regression Models from SAS EM

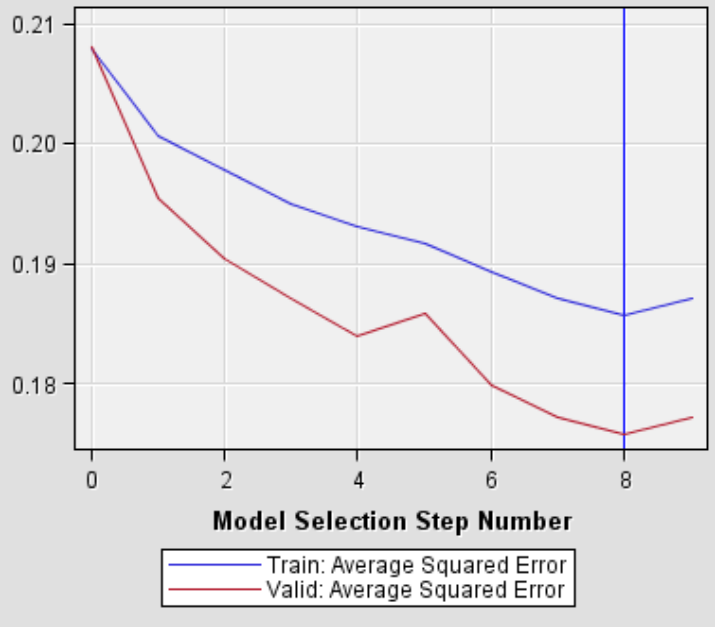


Figure F1. Iteration plot for Percentage Value dataset.

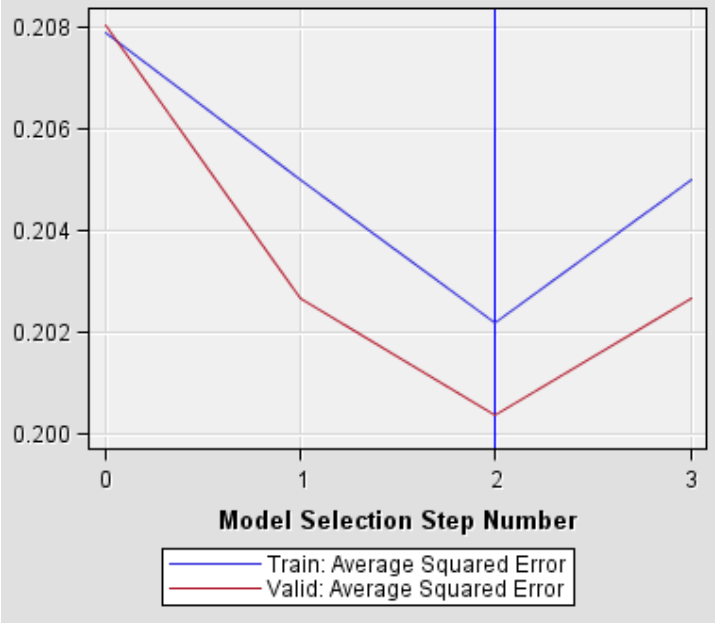


Figure F2. Iteration plot for Monetary Value dataset.

Appendix G

Comparative Cumulative Lift Charts from SAS EM

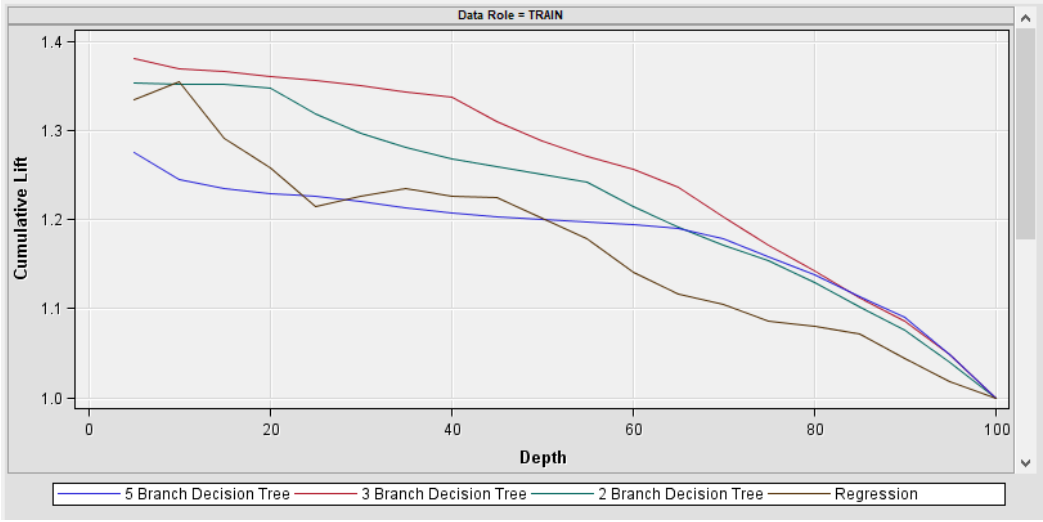


Figure G1. Cumulative lift chart Percentage Value decision tree and regression model comparison.

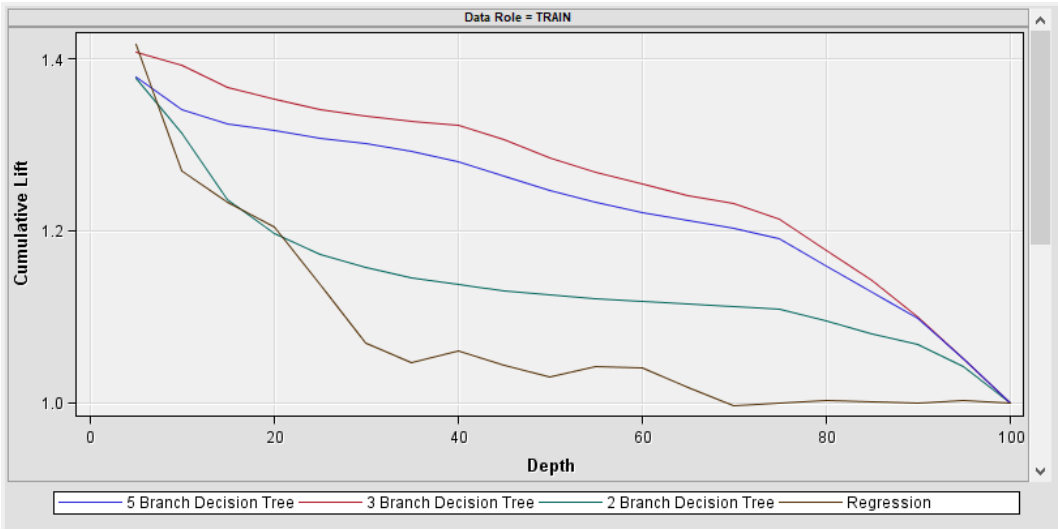


Figure G2. Cumulative lift chart Monetary Value decision tree and regression model comparison.

Appendix H

ROC Charts from SAS EM

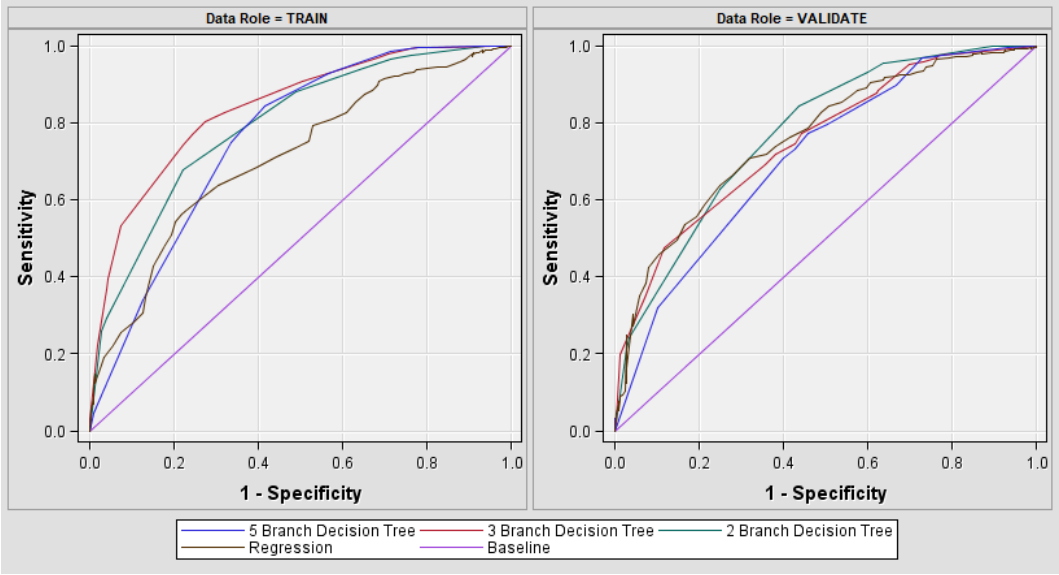


Figure H1. ROC chart for Profit_Loss target variable in Percentage Value dataset.

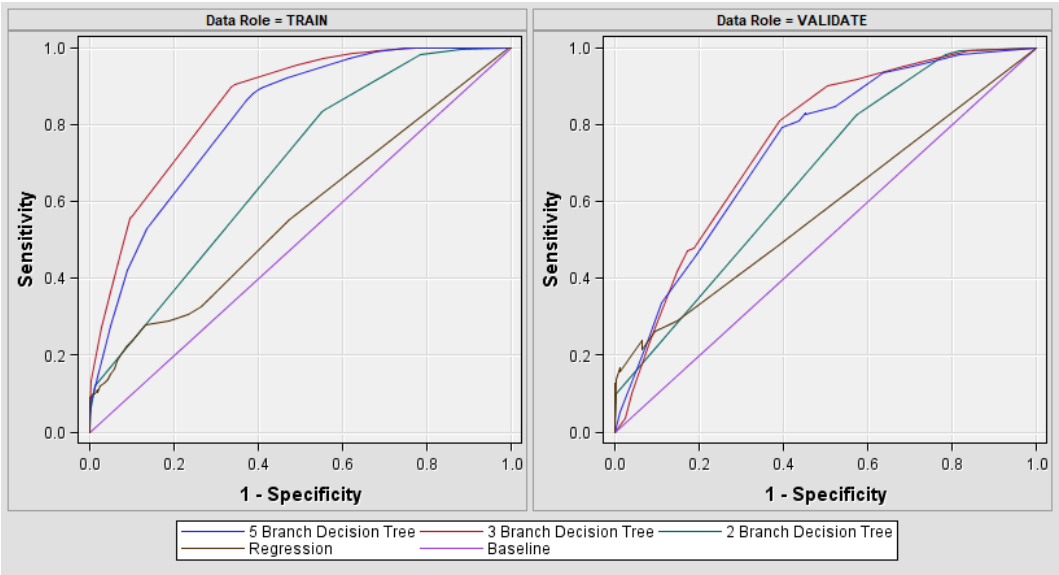


Figure H2. ROC chart for Profit_Loss target variable in Monetary Value dataset**.**