

# Eren Burak Elevli

[eburakelevli@gmail.com](mailto:eburakelevli@gmail.com)

## Table of Contents

<b><i>Profiling and Narrating the Data .....</i></b>	<b><i>2</i></b>
<b><i>Building the Model .....</i></b>	<b><i>7</i></b>
<b><i>1-)Logistic Regression Models .....</i></b>	<b><i>8</i></b>
<b><i>2-)Bayesian Network Models.....</i></b>	<b><i>16</i></b>
<b><i>3-)Decision Tree Models .....</i></b>	<b><i>20</i></b>
<b><i>4-)Forest Models.....</i></b>	<b><i>22</i></b>
<b><i>Pipeline Comparison .....</i></b>	<b><i>24</i></b>
<b><i>Case Study.....</i></b>	<b><i>24</i></b>

# Profiling and Narrating the Data

## Reservation\_status

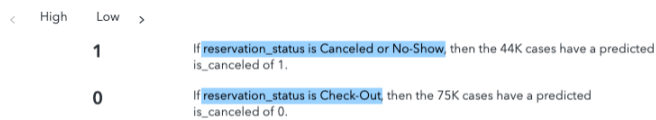
What are the characteristics of is\_canceled?

is\_canceled ranges from 0 to 1. Average is\_canceled is .37. Most cases (96K of 119K) have an is\_canceled between 0 and 1. reservation\_status best differentiates the highest (top 10%) and the lowest (bottom 10%) is\_canceled cases.

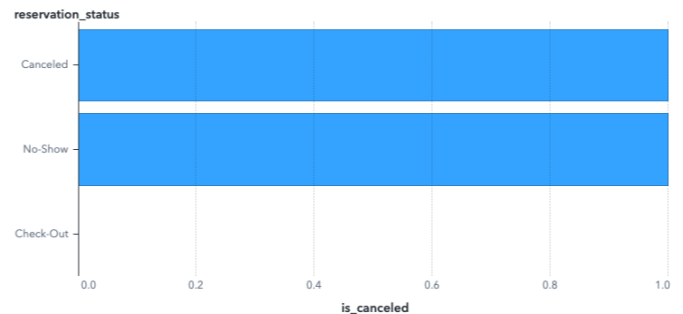
What factors are most related to is\_canceled?



What are the groups based on reservation\_status by the average value of is\_canceled?



What is the relationship between is\_canceled and reservation\_status?



When reservation\_status is Canceled or No-Show, the average of is\_canceled is a high value. When reservation\_status is Check-Out, the average of is\_canceled is a low value. The most common reservation\_status value is Check-Out.

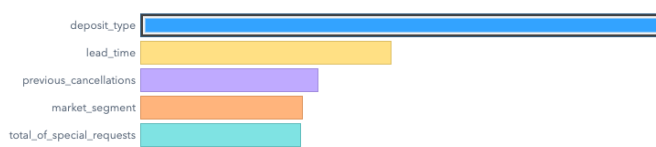
First of all, to have a better understanding of my data and learn the characteristics of it, I have implemented Automated Explanation object. We see here that reservation\_status has the most affect on our target is\_canceled. But when we implement a prediction method for is\_canceled, we can see that reservation\_status can't be used since it shows the result of is\_canceled. So, when building my model I will not include reservation\_status as a feature. In the next chart, I will continue with excluding the reservation\_status.

## Deposit\_type → 1

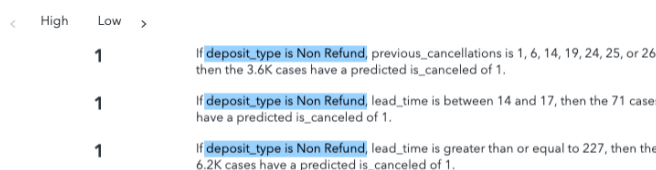
What are the characteristics of is\_canceled?

is\_canceled ranges from 0 to 1. Average is\_canceled is .37. Most cases (96K of 119K) have an is\_canceled between 0 and 1. deposit\_type best differentiates the highest (top 10%) and the lowest (bottom 10%) is\_canceled cases.

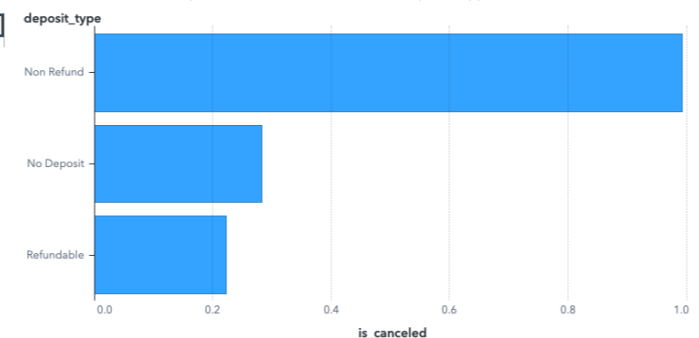
What factors are most related to is\_canceled?



What are the groups based on deposit\_type by the average value of is\_canceled?

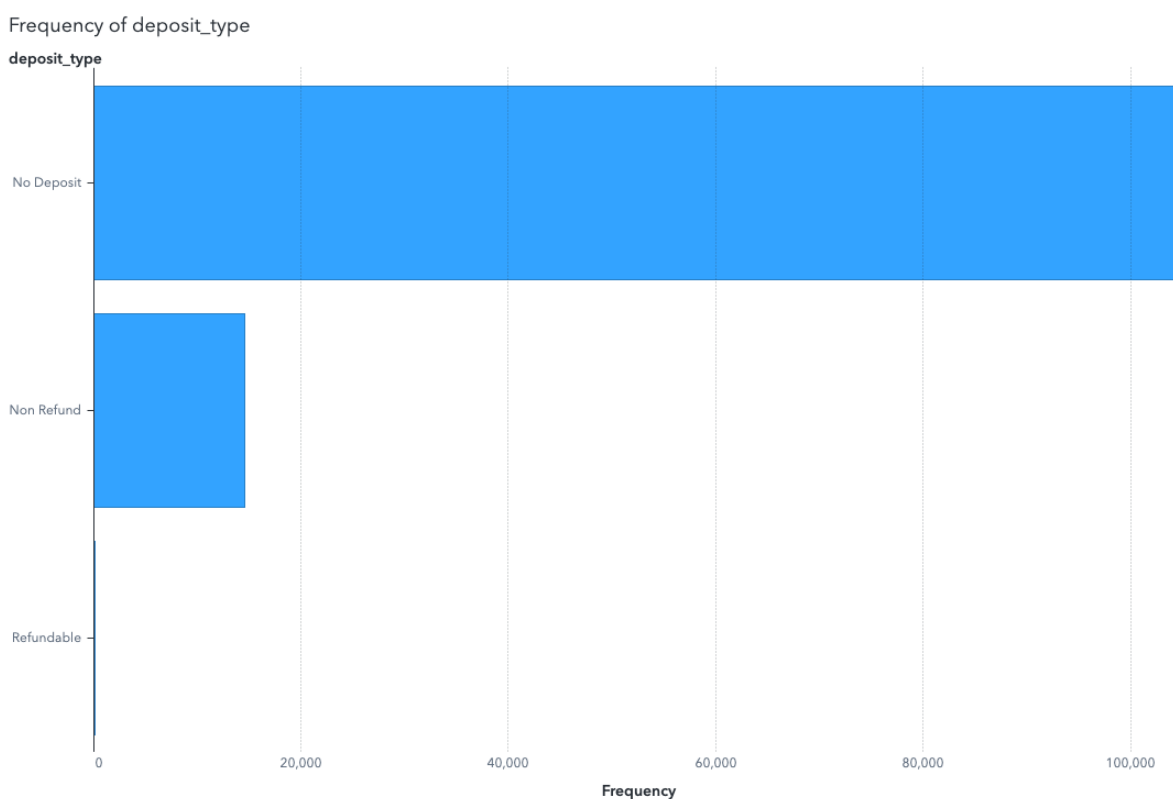


What is the relationship between is\_canceled and deposit\_type?



When deposit\_type is Non Refund, the average of is\_canceled is a high value. When deposit\_type is No Deposit or Refundable, the average of is\_canceled is a low value. The most common deposit\_type value is No Deposit.

We can see here that deposit\_type is the biggest factor to interpret is\_canceled. It is seen that the most cancelations are made in non-refund type which seems counter intuitive but as you can see in the frequency of deposit\_type at the table below that Non Refund is only made by 14,587 people.



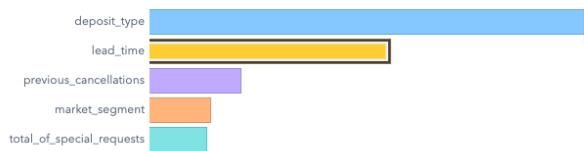
The most deposit\_type made is No Deposit by 104,641 people and Refundable made by 162 people.

Lead\_time → 0.4734

What are the characteristics of is\_canceled?

is\_canceled ranges from 0 to 1. Average is\_canceled is .37. Most cases (96K of 119K) have an is\_canceled between 0 and 1. deposit\_type best differentiates the highest (top 10%) and the lowest (bottom 10%) is\_canceled cases.

What factors are most related to is\_canceled?



What are the groups based on lead\_time by the average value of is\_canceled?

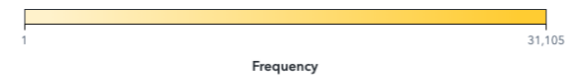
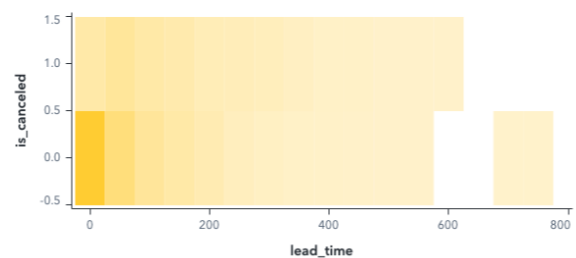
	<	High	Low	>
1				
1				
1				

If hotel is City Hotel, lead\_time is greater than or equal to 521, then the 342 cases have a predicted is\_canceled of 1.

If distribution\_channel is TA/TO, lead\_time is greater than or equal to 559, then the 247 cases have a predicted is\_canceled of 1.

If booking\_changes is less than 1, lead\_time is greater than or equal to 559, then the 247 cases have a predicted is\_canceled of 1.

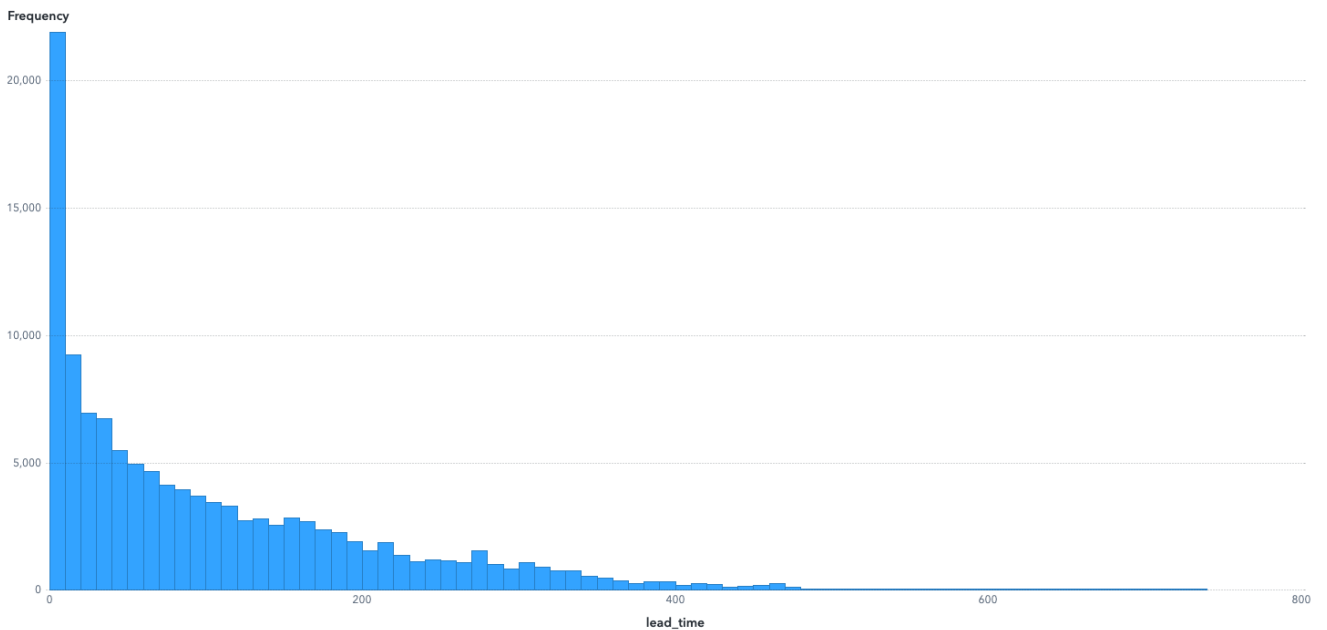
What is the relationship between is\_canceled and lead\_time?



is\_canceled may have a weak relationship with lead\_time. Average lead\_time is 104, and it ranges from 0 to 737.

When we look at the correlation matrix of lead\_time it seems that the smaller the lead\_time is, more people is\_canceled is 1. But when we also investigate the frequency table of lead\_time as seen above we can see that the smaller the lead\_time the more reservation people will make.

Frequency of lead\_time



Because investigating an interval such as lead\_time 0-10 (21,876 people) compare to lead\_time 340-350(539 people) contains more people, it can be more likely to find

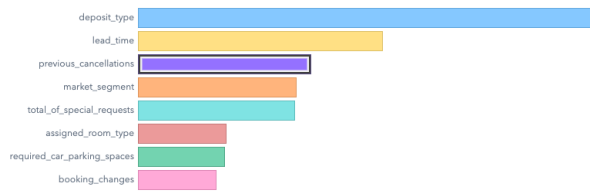
is\_canceled=1 more. So, because lead\_time confusion matrix doesn't tell us anything different than that, lead\_time and is\_canceled may have a weak relationship.

Previous\_cancellations → 0.3348

What are the characteristics of is\_canceled?

is\_canceled ranges from 0 to 1. Average is\_canceled is .37. Most cases (96K of 119K) have an is\_canceled between 0 and 1. deposit\_type best differentiates the highest (top 10%) and the lowest (bottom 10%) is\_canceled cases.

What factors are most related to is\_canceled?



What are the groups based on previous\_cancellations by the average value of is\_canceled?

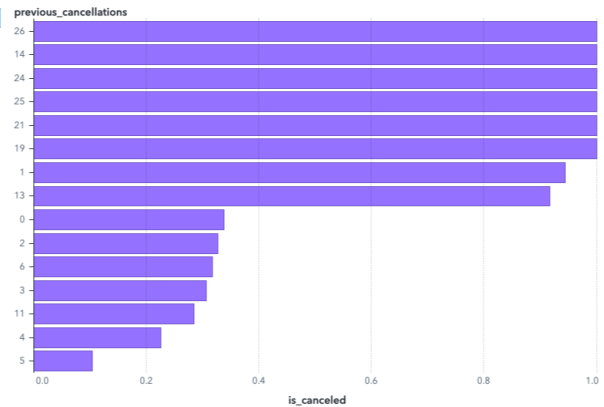
<	High	Low	>
1			
1			
1			

If hotel is Resort Hotel, previous\_cancellations is 14, 19, 24, 25, or 26, then the 132 cases have a predicted is\_canceled of 1.

If distribution\_channel is TA/TO, previous\_cancellations is 13, 19, 21, 24, or 26, then the 105 cases have a predicted is\_canceled of 1.

If booking\_changes is less than 1, previous\_cancellations is 13, 14, 19, 21, 24, 25, or 26, then the 144 cases have a predicted is\_canceled of 1.

What is the relationship between is\_canceled and previous\_cancellations?



When previous\_cancellations is 26, 14, 24, 25, 21, 19, 1 or 13, the average of is\_canceled is a high value. When previous\_cancellations is 0, 2, 6, 3, 11, 4 or 5, the average of is\_canceled is a low value. The most common previous\_cancellations value is 0.0.

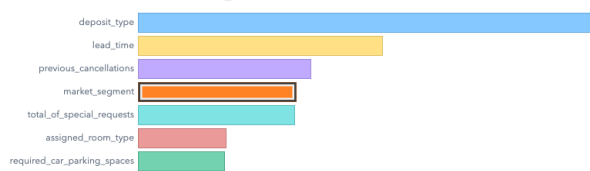
We can see here that, the higher cancellations that customers made before the higher they will cancel their future hotel bookings.

Market\_segment → 0.3061

What are the characteristics of is\_canceled?

is\_canceled ranges from 0 to 1. Average is\_canceled is .37. Most cases (96K of 119K) have an is\_canceled between 0 and 1. deposit\_type best differentiates the highest (top 10%) and the lowest (bottom 10%) is\_canceled cases.

What factors are most related to is\_canceled?



What are the groups based on market\_segment by the average value of is\_canceled?

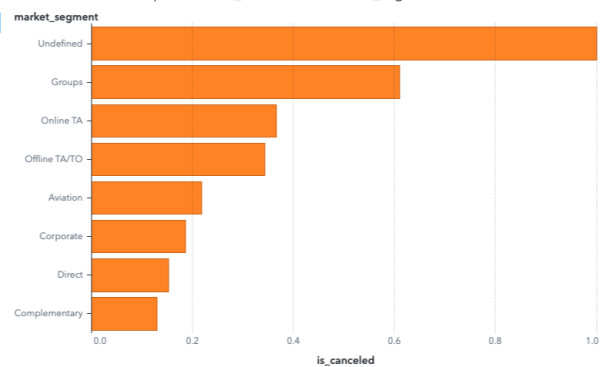
<	High	Low	>
1			
1			
1			

If market\_segment is Groups, lead\_time is greater than or equal to 559, then the 247 cases have a predicted is\_canceled of 1.

If previous\_cancellations is 1, 13, 14, 19, 21, 24, 25, or 26, market\_segment is Aviation or Groups, then the 3.3K cases have a predicted is\_canceled of 1.

If deposit\_type is Non Refund, market\_segment is Offline TA/TO, then the 5K cases have a predicted is\_canceled of 1.

What is the relationship between is\_canceled and market\_segment?



When market\_segment is Undefined, the average of is\_canceled is a high value. When market\_segment is Online TA, Offline TA/TO, Aviation, Corporate, Direct or Complementary, the average of is\_canceled is a low value. The most common market\_segment value is Online TA.

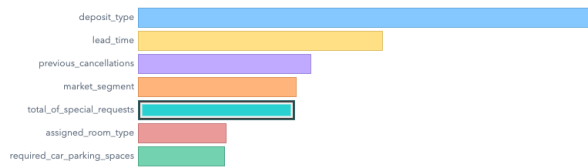
It is seen that if the market\_segment is undefined that is is likely that is\_canceled=1. If the market\_segment is Groups then it is around 0.6 that is\_canceled=1. The others segments have a low value of is\_canceled.

Total\_of\_special\_requests → 0.3023

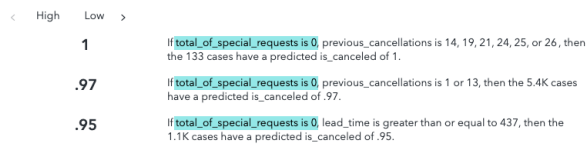
What are the characteristics of is\_canceled?

is\_canceled ranges from 0 to 1. Average is\_canceled is .37. Most cases (96K of 119K) have an is\_canceled between 0 and 1. deposit\_type best differentiates the highest (top 10%) and the lowest (bottom 10%) is\_canceled cases.

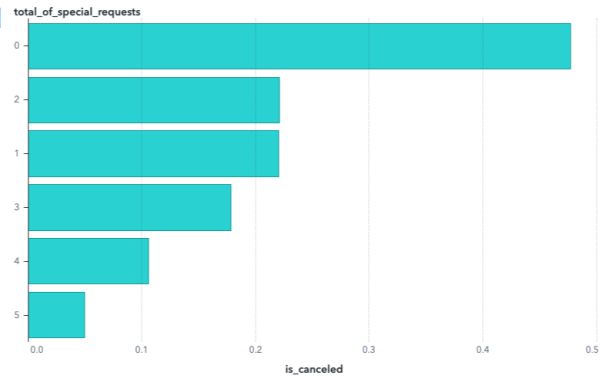
What factors are most related to is\_canceled?



What are the groups based on total\_of\_special\_requests by the average value of is\_canceled?



What is the relationship between is\_canceled and total\_of\_special\_requests?

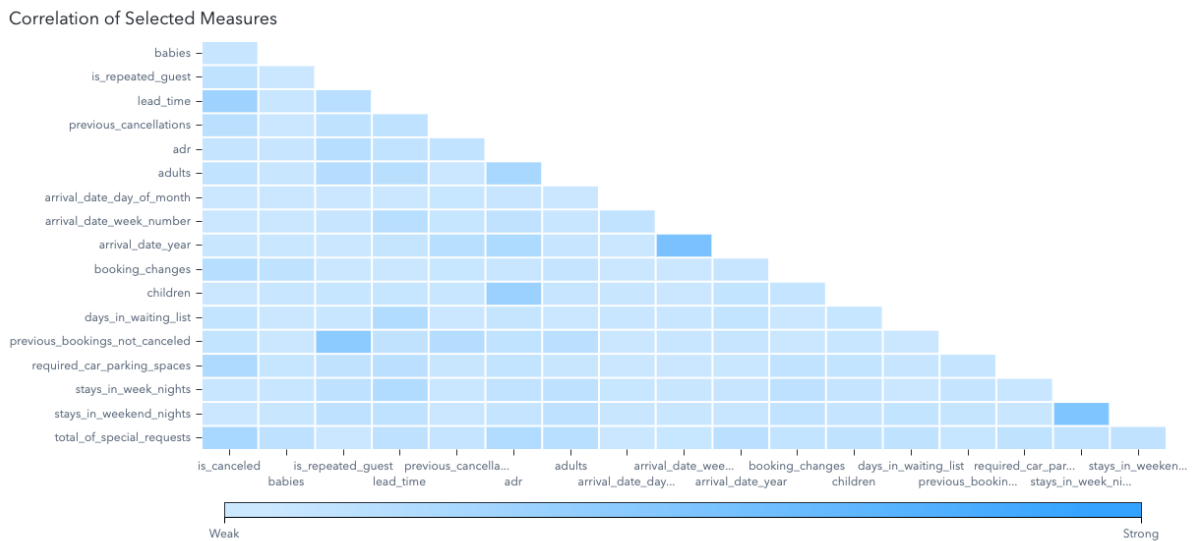


When total\_of\_special\_requests is 0, the average of is\_canceled is a high value. When total\_of\_special\_requests is 3, 4 or 5, the average of is\_canceled is a low value. The most common total\_of\_special\_requests value is 0.0.

total\_of\_special\_requests doesn't seem very relevant with is\_canceled. The highest is\_canceled average is around 0.4 for 0 special request which is quite low.

After eliminating reservation\_status the most related factor is deposit\_type. It is ranked as 1.0. After that, lead\_time has a relative importance of 0.4734 which means it is 0.47 times as important as deposit\_type. The other relative importance are as follows; Previous\_cancellations 0.3348, Market\_segment 0.3061, Total\_of\_special\_requests 0.3023. I will stop investigating from relevant to irrelevant with Total\_of\_special\_requests since the factor after that have a really small relative relevance coefficients.

## Correlation



Is\_repeated\_guest and previous\_bookings\_not\_cancelled has a moderate relations of 0.4181. stays\_in\_weekend\_nights and stays\_in\_week\_nights also shows 0.4990 moderate relation. These features seems to be the only noticeable relevant ones which each other.

## Building the Model

Our target is to interpret is\_canceled and is\_canceled is a binary variable. So, I will use models that can predict binary variables. Using linear regression for interpreting binary variables is against the definition of linear regression so, I will not use linear regression. On the other hand, logistic regression is what should be used in this case because it is formed to predict the binary variables. Then I will continue with using the following methods; Bayesian network, decision tree and forest.

When starting the build the model from hotel data, I select the target variable as is\_canceled. It's level is binary and the order is default. Also as I mentioned in the Profiling and Narrating the Data selection reservation\_status shouldn't be put into the model and since reservation\_status\_date is directly correlated with reservation\_status I also should not include it in my model. So I convert their roles to rejected so that my model doesn't include it.

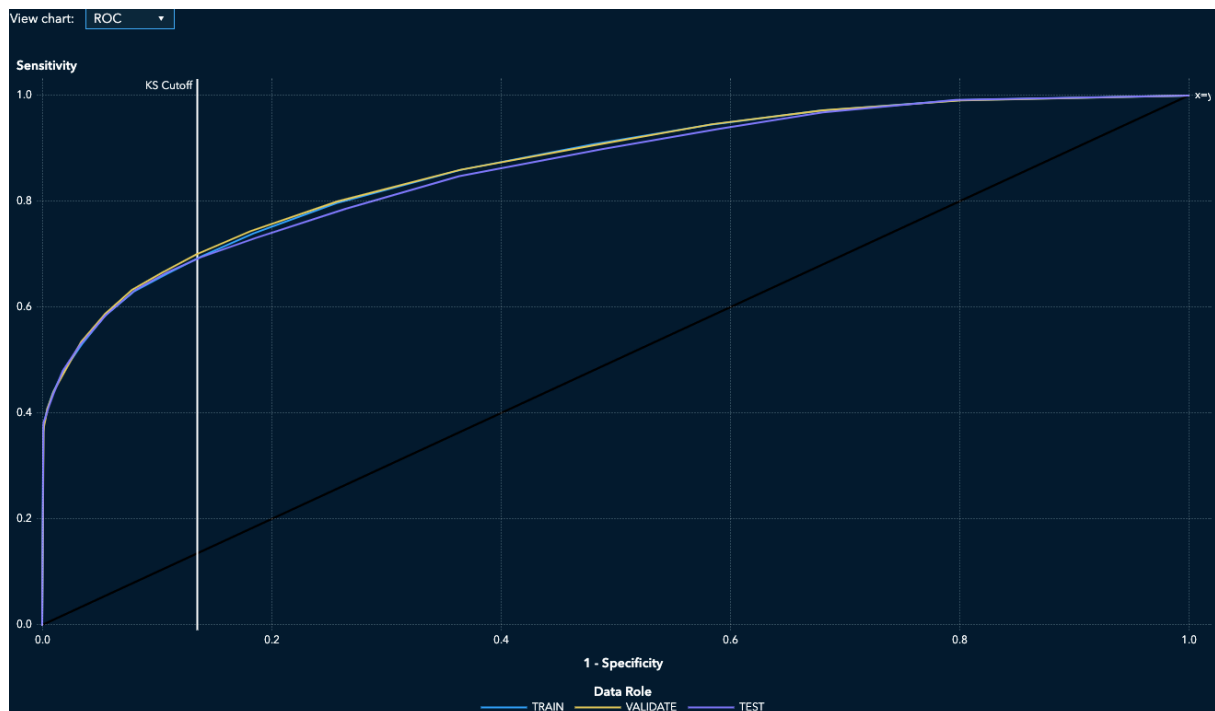
## 1-) Logistic Regression Models

Following model's (a,b,c,d,e) settings only differ in the selection method. I have trained the methods in selection methods None, Stepwise, Backward, Forward and Fastback. Their effect-selection criterion is significance level, selection-process stopping criterion is significance level. I have select these hyperparameters as significance level since I have taken a statistical course which we have deeply discussed this topic. Significance level is a hypothesis testing method. We compare our null hypothesis with an alternative hypothesis and act accordingly while building our model. The model-selection criterion is validation ASE which considers the average square errors and select's the model with the smallest validation ASE.

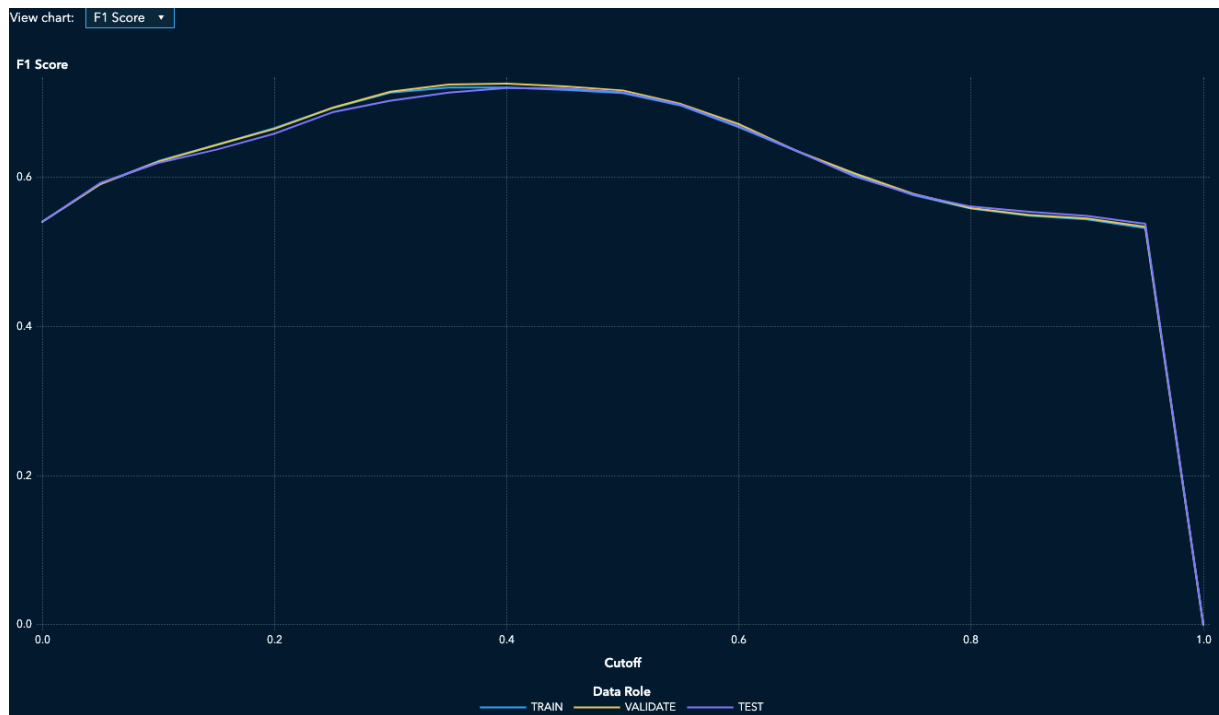
### a) Selection Method: None

Average Square Error for Validation = 0.1323

There is no selection method for this model.



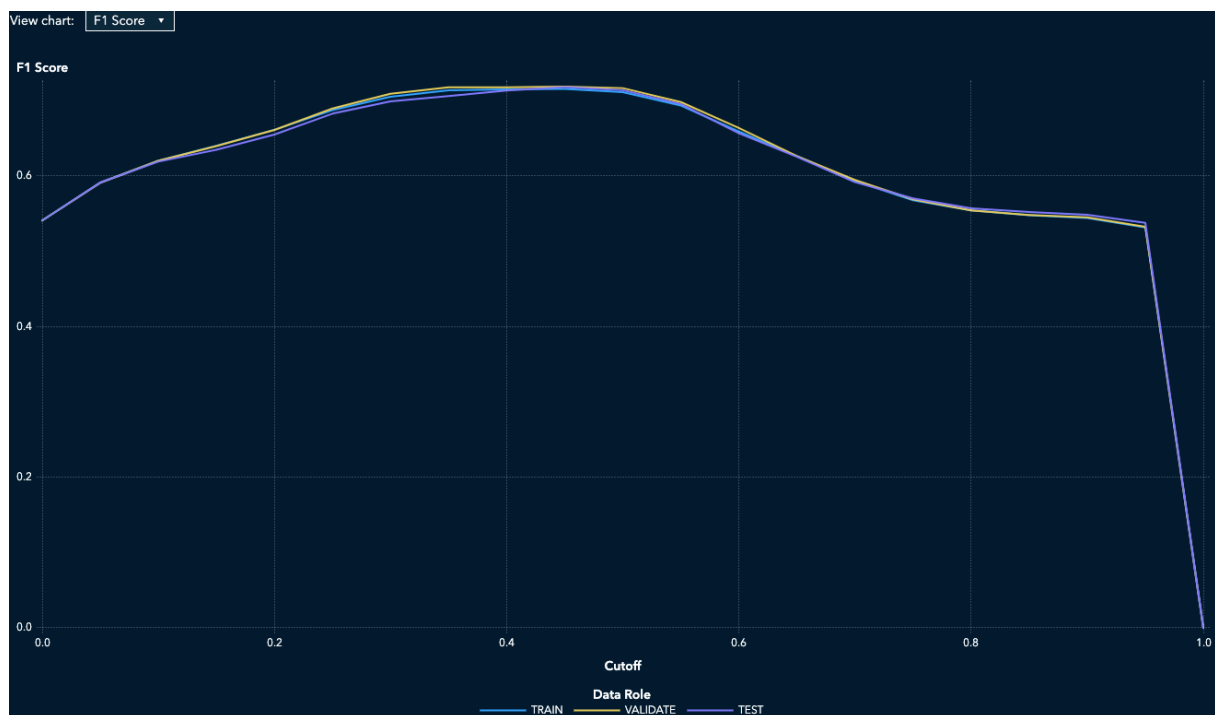
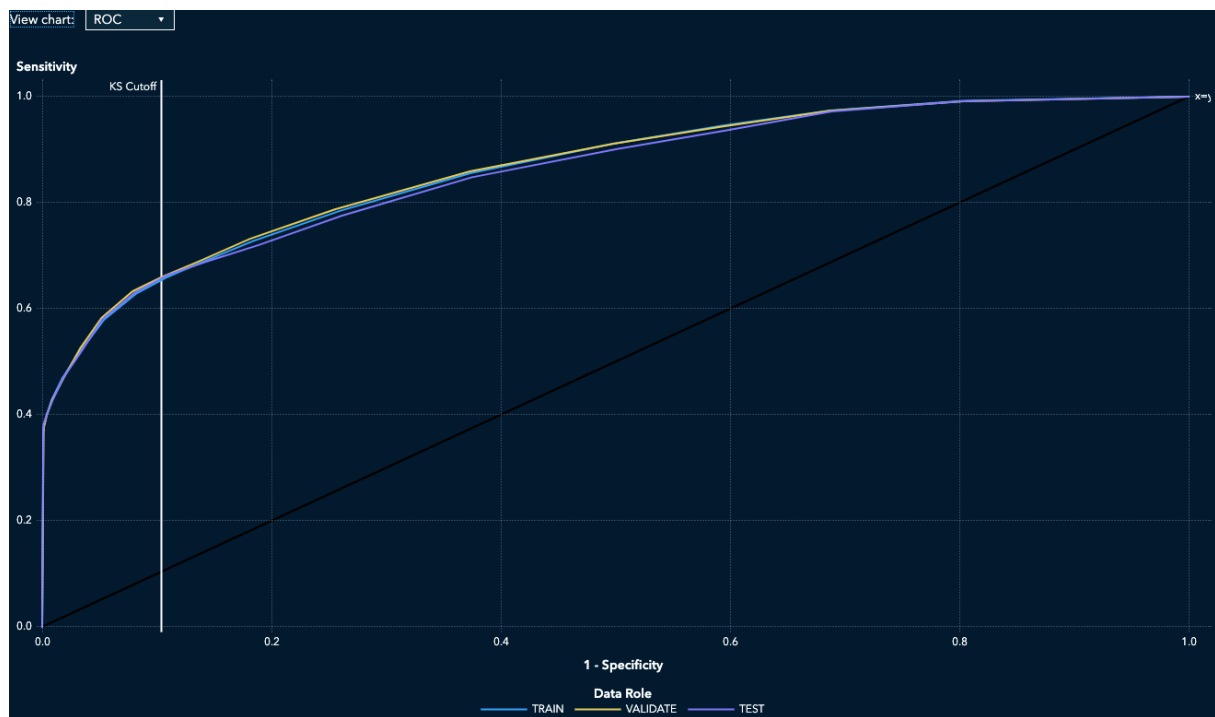




## b) Selection Method:Stepwise

Average Square Error for Validation = 0.1338

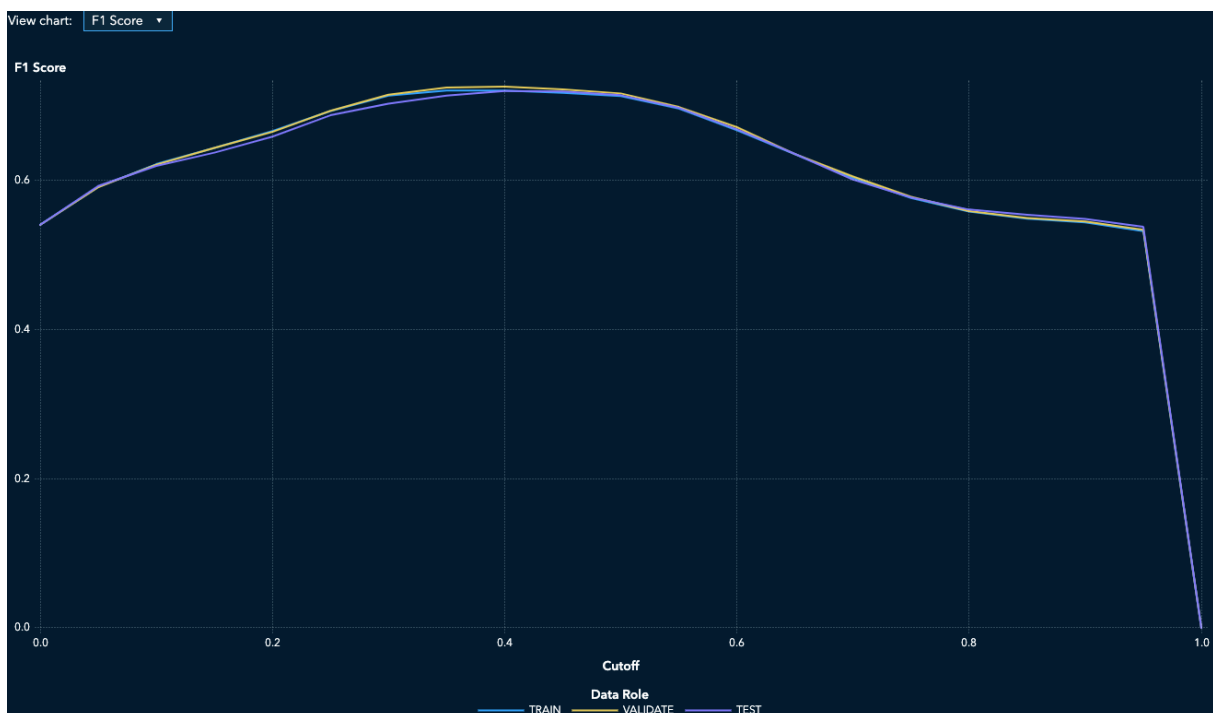
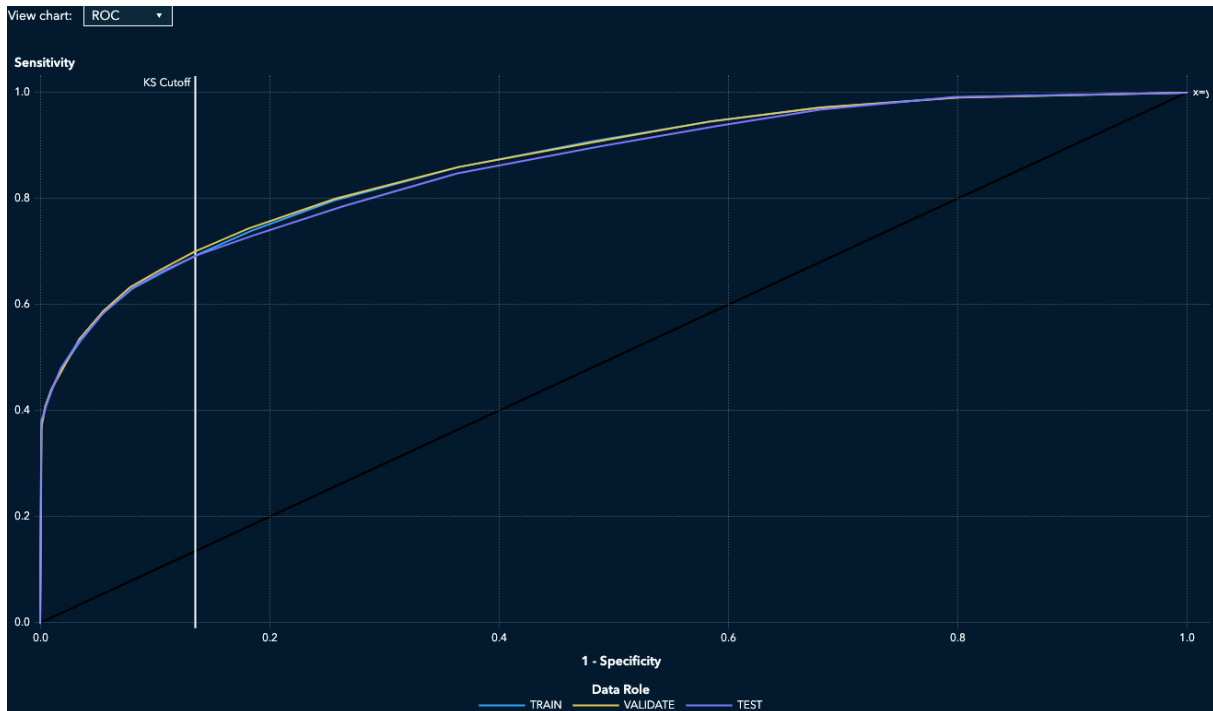
This method begins with considering none of the features, then the most significant feature is added one by one and also along the way features can be removed, until stopping criterion significance level is hit or there is no feature left to consider.



### c) Selection Method: Backward

Average Square Error for Validation = 0.1323

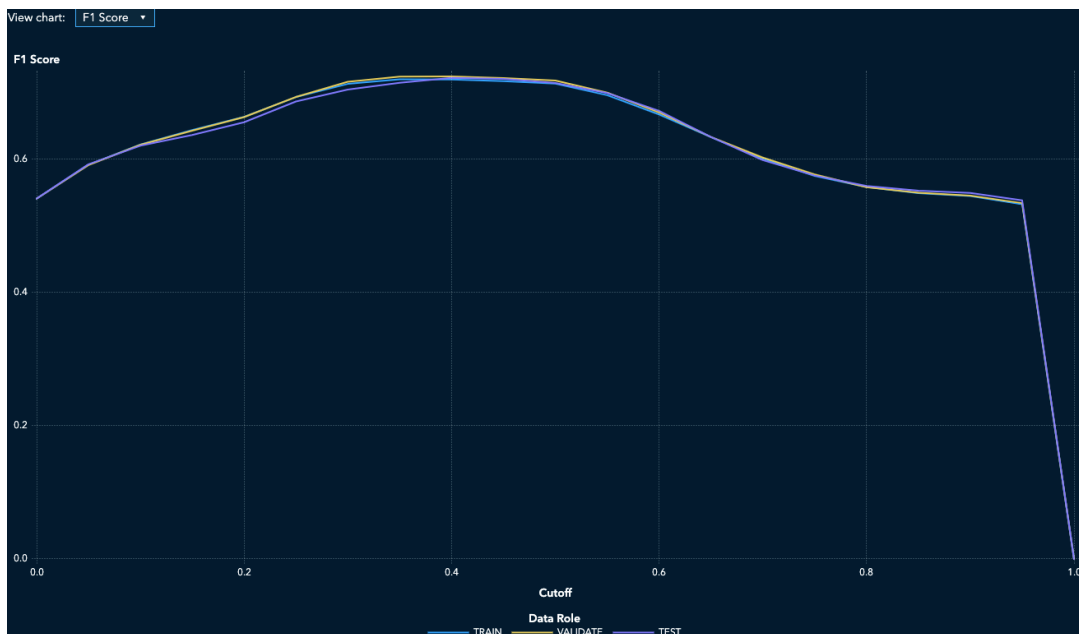
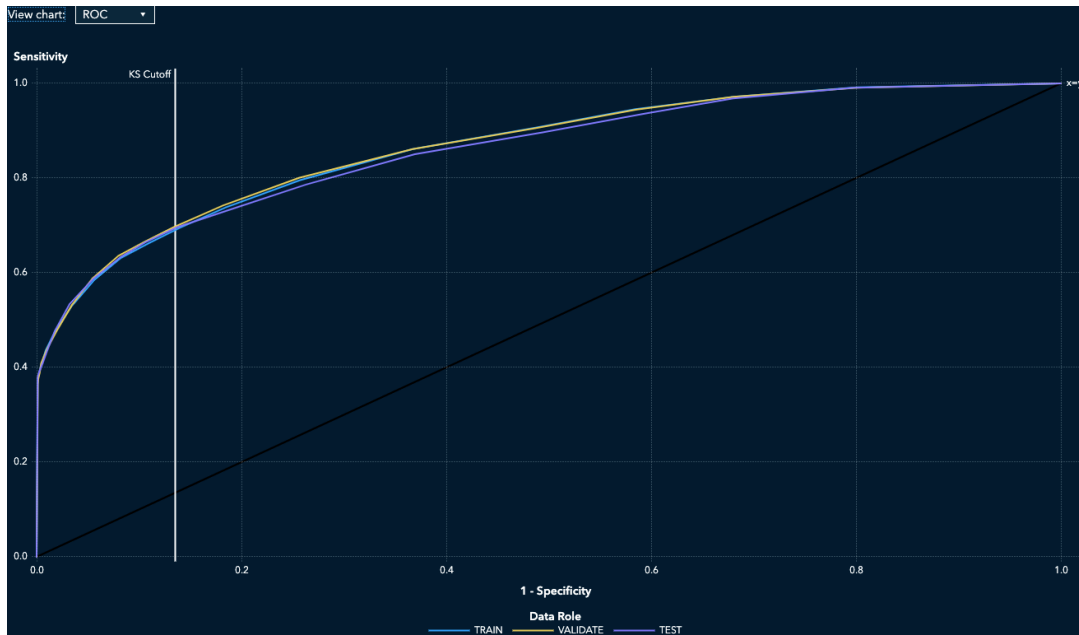
This method begins with considering all features, then the least significant feature is eliminated one by one, until stopping criterion significance level is hit or there is no feature left to consider.



#### d) Selection Method:Forward

Average Square Error for Validation = 0.1332

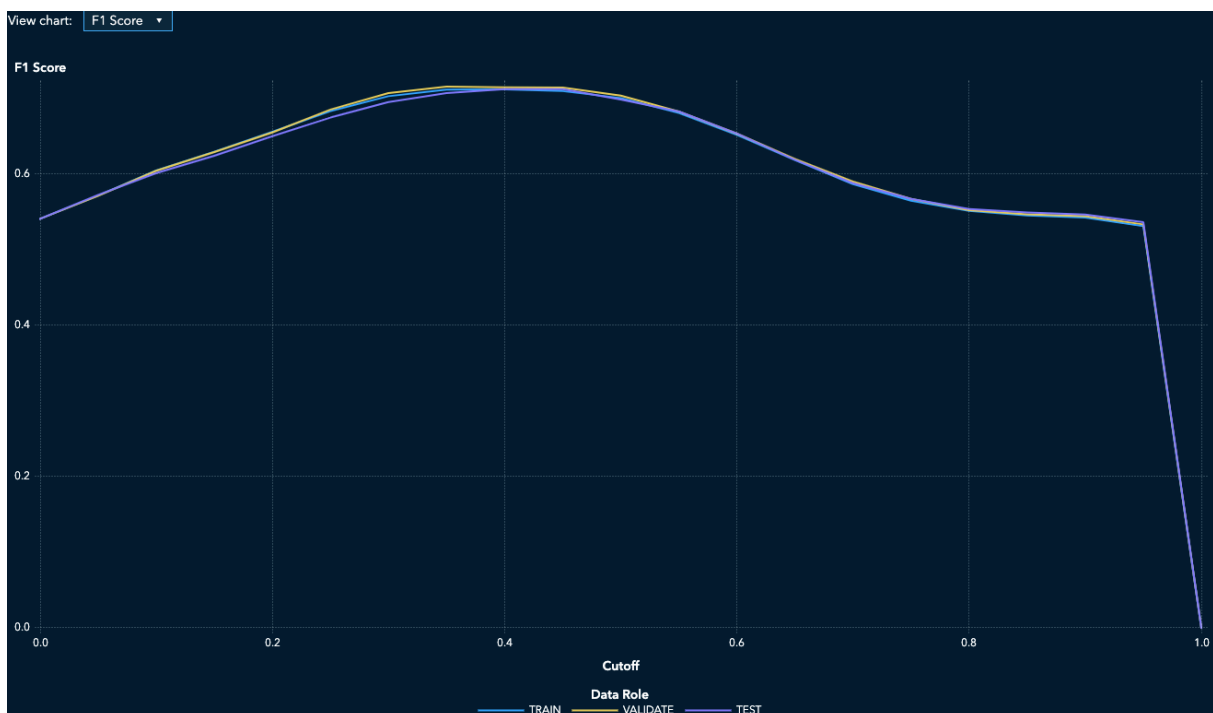
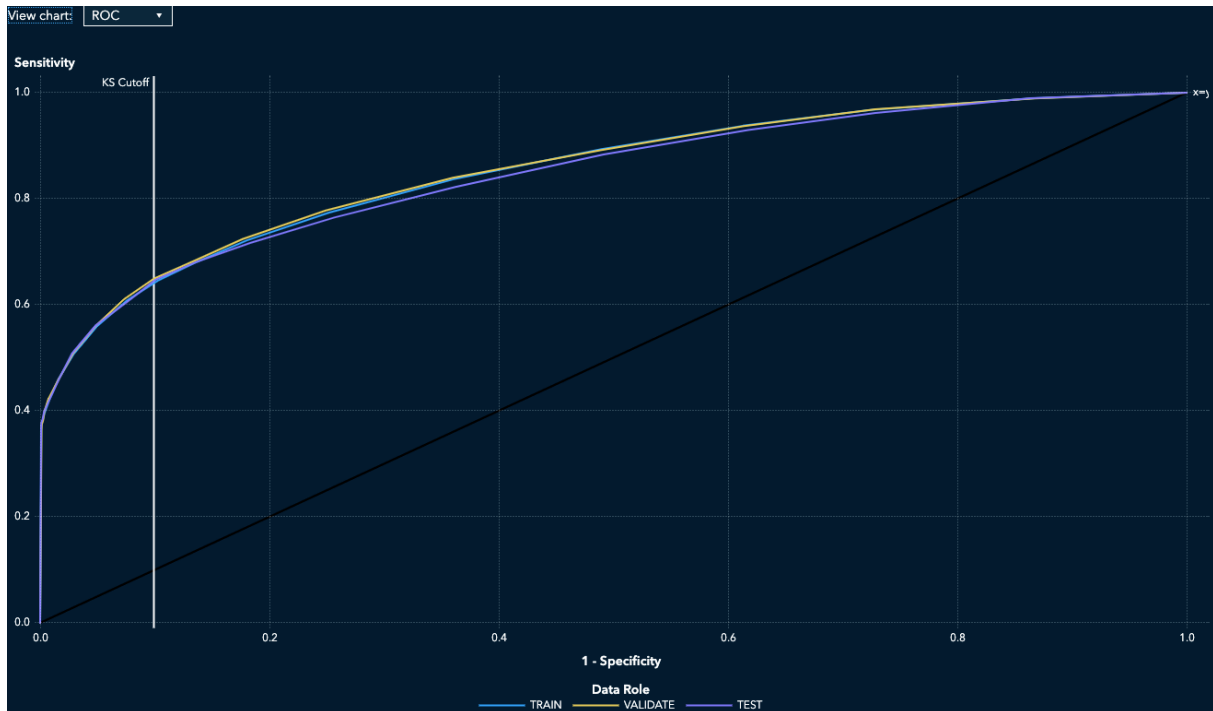
This method begins with considering none of the features, then the most significant feature is added one by one, until stopping criterion significance level is hit or there is no feature left to consider.



### e) Selection Method:Fastback

Average Square Error for Validation = 0.1377

This method begins with considering all features, then the least significant feature is eliminated one by one and without refitting the model, until stopping criterion significance level is hit or there is no feature left to consider.



## Model Comparison

F1 and ROC curve seems pretty close to each other, so in order to see the exact numbers and compare them, I have runned the following model comparisons.

Model Comparison				
Champion	Name	Algorithm Name	Area Under ROC	Misclassification Rate
□	Lo_Reg_No_Sel_Base	Logistic Regression	0.8605	0.1872
	Lo_Reg_Backward	Logistic Regression	0.8605	0.1872
	Lo_Reg_FastBack	Logistic Regression	0.8605	0.1872
	Lo_Reg_Forward	Logistic Regression	0.8601	0.1875
	Lo_Reg_Stepwise	Logistic Regression	0.8570	0.1879

Model Comparison				
Champion	Name	Algorithm Name	F1 Score	Misclassification Rate
□	Lo_Reg_No_Sel_Base	Logistic Regression	0.7143	0.1872
	Lo_Reg_Backward	Logistic Regression	0.7143	0.1872
	Lo_Reg_FastBack	Logistic Regression	0.7143	0.1872
	Lo_Reg_Forward	Logistic Regression	0.7140	0.1875
	Lo_Reg_Stepwise	Logistic Regression	0.7135	0.1879

When we compare the ROC and F1 scores No Selection method used, Backward selection and Fastback selection method's scores are equal but SAS select's Backward selection model as the best model.

So, I will try to improve the model with the Backward selection further and also this results seems perfect to use ensemble method since the ROC and F1 scores are same and ensembling combines the posterior probabilities or predicted values. So, there is a chance of improvement.

### f) Selection Method :Backward (Improved)

All the settings are default as the previous models except effect option and elimination method. Elimination method is selected as backward and I have added an polynomial effect option with a degree of 2.

## Ensemble

Lastly to combine the posterior probabilities or predicted values, I have introduced a postprocessing node, ensemble. I have connected ensemble with a-),c-),e-),f-)

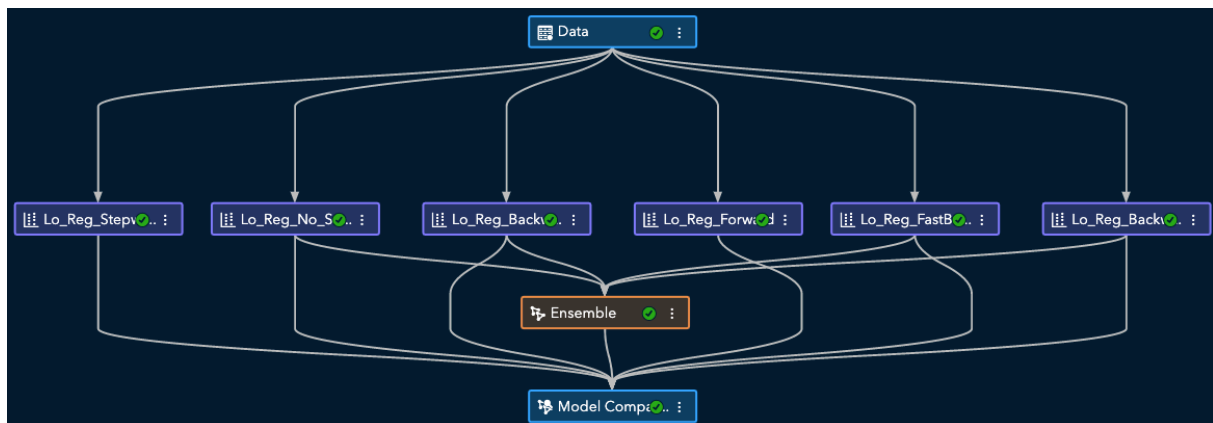
## Conclusion

Model Comparison				
Champion	Name	Algorithm Name	Area Under ROC	Misclassification Rate
🏆	Lo_Reg_Backward_Improved	Logistic Regression	0.8643	0.1851
	Ensemble	Ensemble	0.8622	0.1856
	Lo_Reg_No_Sel_Base	Logistic Regression	0.8605	0.1872
	Lo_Reg_Backward	Logistic Regression	0.8605	0.1872
	Lo_Reg_FastBack	Logistic Regression	0.8605	0.1872
	Lo_Reg_Forward	Logistic Regression	0.8601	0.1875
	Lo_Reg_Stepwise	Logistic Regression	0.8570	0.1879

Model Comparison				
Champion	Name	Algorithm Name	F1 Score	Misclassification Rate
🏆	Lo_Reg_Backward_Improved	Logistic Regression	0.7156	0.1851
	Ensemble	Ensemble	0.7155	0.1856
	Lo_Reg_No_Sel_Base	Logistic Regression	0.7143	0.1872
	Lo_Reg_Backward	Logistic Regression	0.7143	0.1872
	Lo_Reg_FastBack	Logistic Regression	0.7143	0.1872
	Lo_Reg_Forward	Logistic Regression	0.7140	0.1875
	Lo_Reg_Stepwise	Logistic Regression	0.7135	0.1879

Ensemble seems to perform better than a-),c-) and e-). But the best model in comparing ROC and F1 Score is the model f-) which is selection method backward with polynomial effect option with a degree of 2.

## Visual Illustration



## 2-)Bayesian Network Models

### a-)Network Structure : Naïve

Since the Bayesian network we learned was Naïve Bayes, I have run my data with the default settings of naïve.

Variable Selection								
Variable	Selected	Chi-Square	Pr > ChiSq	G-Square	Pr > GSq	Mutual Information	DF	Conditional Variables
arrival_date_month	Yes	338.98674	<.0001	342.75674	<.0001	0.06909	11	
arrival_date_year	Yes	50.36447	<.0001	50.25844	<.0001	0.02648	2	
assigned_room_type	Yes	2964.59241	<.0001	3137.59411	<.0001	0.20702	11	
babies	Yes	88.77740	<.0001	99.86327	<.0001	0.03732	4	
children	Yes	53.26147	<.0001	54.16640	<.0001	0.02749	4	
customer_type	Yes	1313.28584	<.0001	1380.31464	<.0001	0.13815	3	
deposit_type	Yes	16561	<.0001	18718	<.0001	0.47954	2	
distribution_channel	Yes	2228.77403	<.0001	2429.04837	<.0001	0.18260	4	
hotel	Yes	1301.98243	<.0001	1332.42034	<.0001	0.13575	1	
is_repeated_guest	Yes	512.77652	<.0001	588.31066	<.0001	0.09044	1	
market_segment	Yes	5172.01664	<.0001	5323.85856	<.0001	0.26763	6	
meal	Yes	203.58108	<.0001	201.83936	<.0001	0.05304	4	
previous_cancellations	Yes	5549.86568	<.0001	5864.44773	<.0001	0.28037	13	
required_car_parking_spaces	Yes	2783.55170	<.0001	4277.87139	<.0001	0.24077	3	
reserved_room_type	Yes	406.57607	<.0001	416.17166	<.0001	0.07611	9	
stays_in_weekend_nights	Yes	75.08015	<.0001	74.78109	<.0001	0.03230	14	
total_of_special_requests	Yes	4952.50383	<.0001	5150.42193	<.0001	0.26339	5	
adr	Yes	814.46459	<.0001	889.66767	<.0001	0.11110	9	
adults	No	1.70001	0.6369	1.98652	0.5752	0.00527	3	
arrival_date_day_of_month	Yes	54.32069	<.0001	54.15296	<.0001	0.02749	9	
arrival_date_week_number	Yes	297.39637	<.0001	299.98415	<.0001	0.06465	9	
booking_changes	Yes	161.24767	<.0001	182.89089	<.0001	0.05050	9	
days_in_waiting_list	Yes	300.95248	<.0001	289.26627	<.0001	0.06348	8	
lead_time	Yes	5823.88643	<.0001	5881.41772	<.0001	0.28076	9	
previous_bookings_not_canceled	Yes	130.87182	<.0001	171.94528	<.0001	0.04896	9	
stays_in_week_nights	Yes	71.78946	<.0001	73.37543	<.0001	0.03200	9	

The selection of the variables seems consistent with the independence test. Adults were not included to the model since it fails the independence test and has a high value compare to other  $Pr > ChiSq \rightarrow 0.6369$  and  $Pr > GSq \rightarrow 0.5752$



## b-)Network Structure : Naïve (Auto-Tune)

To further improve my model, I have implemented a Auto-tuned model with whose network structure is Naïve and all the other settings are default.

Variable Selection								
Variable	Selected	Chi-Square	Pr > ChiSq	G-Square	Pr > GSq	Mutual Information	DF	Conditional Variables
arrival_date_month	Yes	338.98674	<.0001	342.75674	<.0001	0.06909	11	
arrival_date_year	Yes	50.36447	<.0001	50.25844	<.0001	0.02648	2	
assigned_room_type	Yes	2964.59241	<.0001	3137.59411	<.0001	0.20702	11	
babies	Yes	88.77740	<.0001	99.86327	<.0001	0.03732	4	
children	Yes	53.26147	<.0001	54.16640	<.0001	0.02749	4	
customer_type	Yes	1313.28584	<.0001	1380.31464	<.0001	0.13815	3	
deposit_type	Yes	16561	<.0001	18718	<.0001	0.47954	2	
distribution_channel	Yes	2228.77403	<.0001	2429.04837	<.0001	0.18260	4	
hotel	Yes	1301.98243	<.0001	1332.42034	<.0001	0.13575	1	
is_repeated_guest	Yes	512.77652	<.0001	588.31066	<.0001	0.09044	1	
market_segment	Yes	5172.01664	<.0001	5323.85856	<.0001	0.26763	6	
meal	Yes	203.58108	<.0001	201.83936	<.0001	0.05304	4	
previous_cancellations	Yes	5549.86568	<.0001	5864.44773	<.0001	0.28037	13	
required_car_parking_spaces	Yes	2783.55170	<.0001	4277.87139	<.0001	0.24077	3	
reserved_room_type	Yes	406.57607	<.0001	416.17166	<.0001	0.07611	9	
stays_in_weekend_nights	Yes	75.08015	<.0001	74.78109	<.0001	0.03230	14	
total_of_special_requests	Yes	4952.50383	<.0001	5150.42193	<.0001	0.26339	5	
adr	Yes	494.38003	<.0001	541.34677	<.0001	0.08677	11	
adults	Yes	10.19963	0.0698	11.91897	0.0359	0.01290	5	
arrival_date_day_of_month	Yes	40.97389	<.0001	40.98802	<.0001	0.02392	11	
arrival_date_week_number	Yes	312.21476	<.0001	315.33497	<.0001	0.06628	11	
booking_changes	Yes	429.68710	<.0001	473.43709	<.0001	0.08116	11	
days_in_waiting_list	Yes	487.06552	<.0001	466.74283	<.0001	0.08059	9	
lead_time	Yes	6211.40045	<.0001	6311.53872	<.0001	0.29041	11	
previous_bookings_not_canceled	Yes	195.78792	<.0001	265.25644	<.0001	0.06080	11	
stays_in_week_nights	Yes	101.44378	<.0001	103.59286	<.0001	0.03801	10	

As we can see here, adults were also selected as a variable.

## c-)Network Structure : Naïve(Auto-Tune / Improved)

To be able to further improve my model I have changed some parameters as, missing class inputs with" selecting impute with mode" . Rather than ignoring a class, this selection replaces the mode of that class who has the missing values.

Also I have changed the missing interval inputs with" selecting impute with mean "options. Rather than ignoring a interval, this selection replaces the mean of that interval who has the missing values.

Rather than ignoring missing class inputs and missing interval inputs, adding them into our model with mode and mean of them, we may see an improvement for our model.

Variable Selection							
Variable	Selected	Chi-Square	Pr > ChiSq	G-Square	Pr > GSq	Mutual Information	DF
arrival_date_month	Yes	339.02615	<.0001	342.79927	<.0001	0.06909	11
arrival_date_year	Yes	50.35198	<.0001	50.24643	<.0001	0.02648	2
assigned_room_type	Yes	2963.59399	<.0001	3136.46890	<.0001	0.20698	11
babies	Yes	88.78453	<.0001	99.87127	<.0001	0.03733	4
children	Yes	53.26546	<.0001	54.17092	<.0001	0.02749	4
customer_type	Yes	1312.59575	<.0001	1379.57163	<.0001	0.13811	3
deposit_type	Yes	16560	<.0001	18718	<.0001	0.47953	2
distribution_channel	Yes	2228.30790	<.0001	2428.26201	<.0001	0.18257	4
hotel	Yes	1302.20990	<.0001	1332.65531	<.0001	0.13576	1
is_repeated_guest	Yes	512.81146	<.0001	588.35044	<.0001	0.09044	1
market_segment	Yes	5173.66575	<.0001	5325.84480	<.0001	0.26768	7
meal	Yes	203.59862	<.0001	201.85885	<.0001	0.05305	4
previous_cancellations	Yes	5549.63726	<.0001	5864.27010	<.0001	0.28036	13
required_car_parking_spaces	Yes	2783.65404	<.0001	4277.99930	<.0001	0.24077	3
reserved_room_type	Yes	406.11709	<.0001	415.68998	<.0001	0.07607	9
stays_in_weekend_nights	Yes	75.10971	<.0001	74.81129	<.0001	0.03231	14
total_of_special_requests	Yes	4951.55886	<.0001	5149.39695	<.0001	0.26337	5
adr	No	0.42399	0.5150	0.42225	0.5158	0.00243	1
adults	Yes	3.39930	0.0652	3.97252	0.0462	0.00745	1
arrival_date_day_of_month	Yes	1.86313	0.1723	1.86305	0.1723	0.00510	1
arrival_date_week_number	No	0.13152	0.7169	0.13152	0.7169	0.00135	1
booking_changes	Yes	2.13666	0.1438	2.41350	0.1203	0.00580	1
days_in_waiting_list	Yes	9.43581	0.0021	9.10576	0.0025	0.01127	1
lead_time	Yes	1530.60735	<.0001	1466.76590	<.0001	0.14236	1
previous_bookings_not_canceled	Yes	9.60737	0.0019	12.08413	0.0005	0.01299	1
stays_in_week_nights	Yes	4.24176	0.0394	5.28247	0.0215	0.00859	1

As we can see in the above table, the changes I made resulted in not selecting the adr and arrival\_date\_week\_number.

## Ensemble

Lastly to combine the posterior probabilities or predicted values, I have introduced a postprocessing node, ensemble. Ensemble creates a new model by using the models I have created above which are Naïve, Naïve(Auto-Tune), Naïve(Auto-Tune / Improved).

## Conclusion

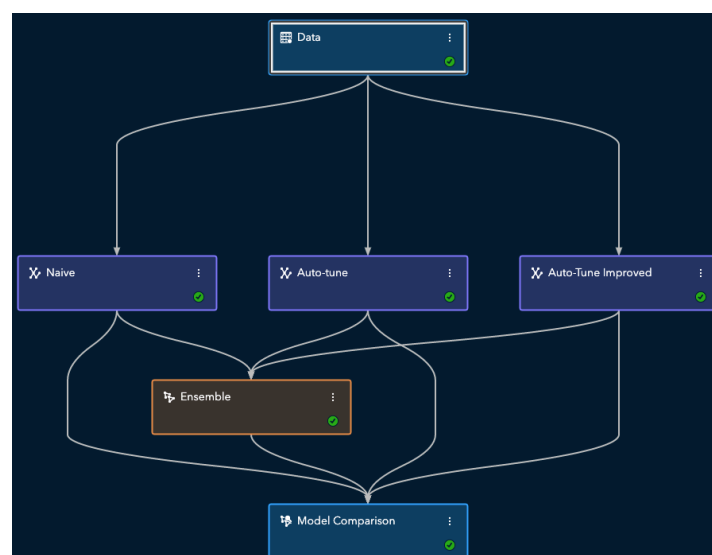
To decide which model is the best, I will compare their ROC and F1 scores.

Model Comparison				
Champion	Name	Algorithm Name	Area Under ROC	Misclassification Rate
	Ensemble	Ensemble	0.8692	0.2092
	Auto-tune	Bayesian Network	0.8641	0.2199
	Auto-Tune Improved	Bayesian Network	0.8604	0.2150
	Naive	Bayesian Network	0.8587	0.2202

Model Comparison				
Champion	Name	Algorithm Name	F1 Score	Misclassification Rate
	Ensemble	Ensemble	0.7250	0.2092
	Auto-tune	Bayesian Network	0.7178	0.2199
	Naive	Bayesian Network	0.7153	0.2202
	Auto-Tune Improved	Bayesian Network	0.7136	0.2150

For ROC scores option b-) seems the best among a-),b-),c-) and the ensemble gave the best ROC score.For F1 Score option b-) seems the best among a-),b-),c-) and the ensemble gave the best F1 score.

## Visual Illustration



### 3-)Decision Tree Models

Selection of splitting methods is important for building our model, since splitting forms the tree. More specifically, splitting is a criteria for going from parent node to it's child nodes.

To determine the best-split method for my data, I build my model only changing the class target criterion. I build three models for to test and find the best-split method.

a-)Class Target Criterion : Information Gain Ratio(Default)

b-)Class Target Criterion : Entropy

c-)Class Target Criterion : Gini

When I compare these models in the sense of ROC and F1 scores, Gini split gave the best result. In order to further improve model c-), I apply a greedy heuristic to my model which is C4.5. Pruning options subtree method is selected as C4.5 .

c-)Class Target Criterion : Gini – C4.5

I again observe an increase in the ROC and F1 scores for this model. To further improve my model again, I increase my models maximum depth to first 20 but SAS couldn't handle 20 depth tree so I change my new models maximum depth to 15. Depth is the most distanced path from the starting point to the deepest leaf in the decision tree. This means more splitting and more features will be included into consideration as the depth of our tree grows. It can also increase our model's ROC and F1 Score as well.

c-)Class Target Criterion : Gini – C4.5 – Depth 15

Once again I managed to improve my model in the sense of ROC and F1 scores.

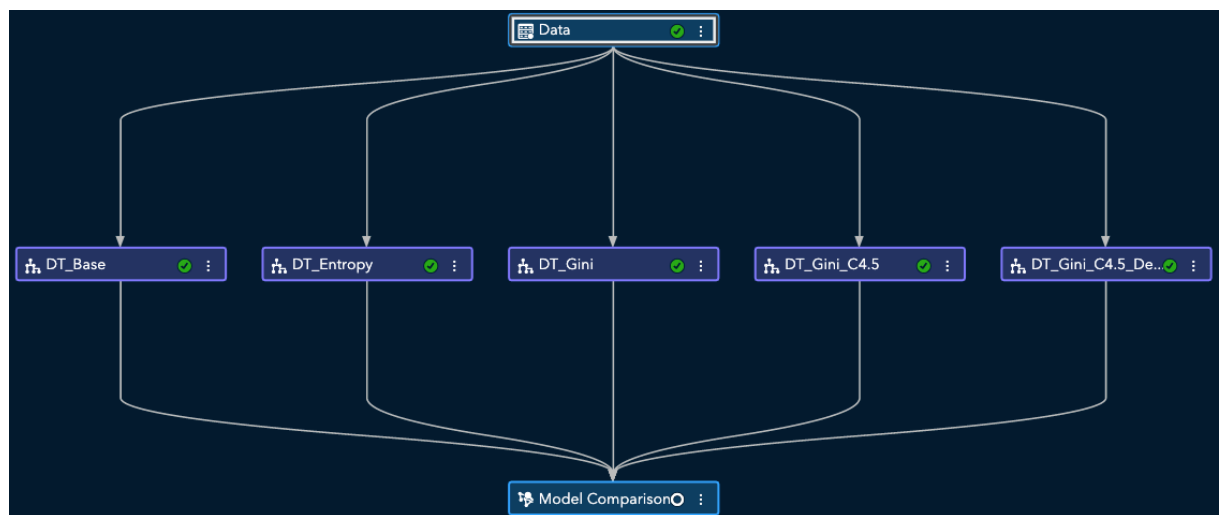
## Conclusion

Model Comparison				
Champion	Name	Algorithm Name	Area Under ROC	Misclassification Rate
🏆	DT_Gini_C4.5_Depth15	Decision Tree	0.8843	0.1765
	DT_Gini_C4.5	Decision Tree	0.8675	0.1806
	DT_Gini	Decision Tree	0.8673	0.1806
	DT_Entropy	Decision Tree	0.8604	0.1840
	DT_Base	Decision Tree	0.7664	0.2301

Model Comparison				
Champion	Name	Algorithm Name	F1 Score	Misclassification Rate
🏆	DT_Gini_C4.5_Depth15	Decision Tree	0.7412	0.1765
	DT_Gini_C4.5	Decision Tree	0.7202	0.1806
	DT_Gini	Decision Tree	0.7201	0.1806
	DT_Entropy	Decision Tree	0.7146	0.1840
	DT_Base	Decision Tree	0.5505	0.2301

As a result decision tree with a Gini split, using a C4.5 heuristic method and with a depth 15 gives us the best model.

## Visual Illustration



## 4-)Forest Models

Selection of splitting methods is important for building our model, since splitting forms the tree and these trees forms our forest model. More specifically, splitting is a criteria for going from parent node to it's child nodes.

To determine the best-split method for my data, I build my model only changing the class target criterion. I build three models for to test and find the best-split method.

a-)Class Target Criterion : Information Gain Ratio(Default)

b-)Class Target Criterion : Entropy

c-)Class Target Criterion : Gini

ROC and F1 scores shows that the Gini is the better split among the model a,b and c. To further improve my model, I will implement some changes

d-)Class Target Criterion :Gini (Depth 50)

Increasing the depth of our model will result in more parent node to child nodes splitting. More splitting means that, model considering more features. This will increase our model's complexity but we can get a better model.So, Depth of the model increased to 50.

e-)Class Target Criterion : Gini (Tree 250)

The more tree, the bigger our forest model will be. The bigger our forest model, it means the more options we consider. There is a possibility of enhancing our model. Number of trees for the model increased to 250.

f-)Class Target Criterion : Gini (Depth 50 & Tree 250)

Depth of the model increased to 50 and the number of trees for the model increased to 250.

## Ensemble


Lastly to combine the posterior probabilities or predicted values, I have introduced 2 postprocessing ensemble nodes.

Ensemble default is the default version.

Ensemble 1's predicted values is changed to maximum, posterior probabilities changed to voting and voting posterior probabilities is average.


## Conclusion

The performance of the model compared with respect to ROC and F1 scores.

Model Comparison				
Champion	Name	Algorithm Name	Area Under ROC	Misclassification Rate
	Forest_Gini_Tree250_Depth50	Forest	0.9203	0.1503
	Forest_Gini_Depth50	Forest	0.9199	0.1487
	Ensemble	Ensemble	0.9194	0.1509
	Ensemble (1)	Ensemble	0.9190	0.1527
	Forest_Gini_Tree250	Forest	0.9186	0.1536
	Forest_Gini (1)	Forest	0.9175	0.1537
	Forest_Entropy	Forest	0.9174	0.1547
	Forest_Base	Forest	0.8844	0.1805

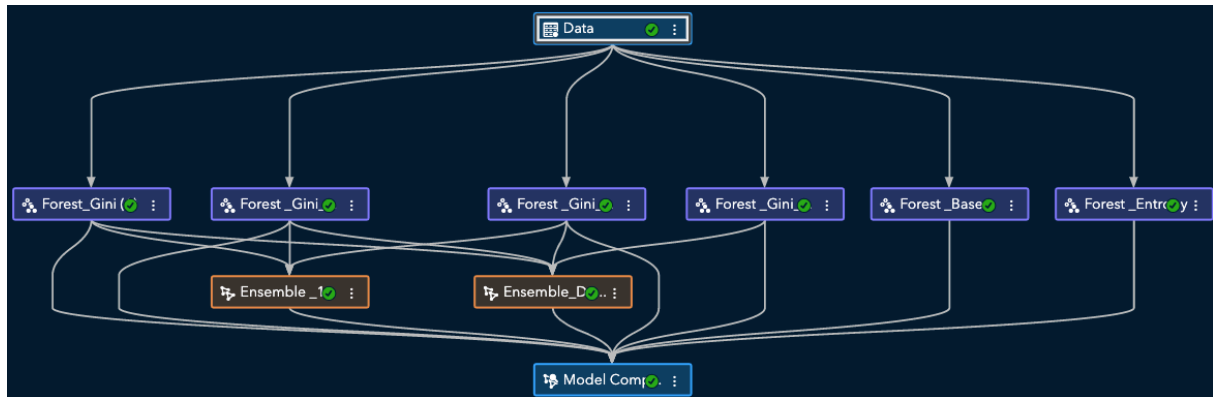
Model Comparison				
Champion	Name	Algorithm Name	F1 Score	Misclassification Rate
	Forest_Gini_Depth50	Forest	0.7802	0.1487
	Forest_Gini_Tree250_Depth50	Forest	0.7775	0.1503
	Ensemble_Default	Ensemble	0.7765	0.1509
	Ensemble_1	Ensemble	0.7733	0.1527
	Forest_Gini_Tree250	Forest	0.7714	0.1536
	Forest_Gini (1)	Forest	0.7712	0.1537
	Forest_Entropy	Forest	0.7686	0.1547
	Forest_Base	Forest	0.7129	0.1805

The winner of the ROC and F1 scores are different, so I also compare them with respect to Accuracy to find the accurate one.

Model Comparison				
Champion	Name	Algorithm Name	Accuracy	Misclassification Rate
	Forest_Gini_Depth50	Forest	0.8513	0.1487
	Forest_Gini_Tree250_Depth50	Forest	0.8497	0.1503
	Ensemble_Default	Ensemble	0.8491	0.1509
	Ensemble_1	Ensemble	0.8473	0.1527
	Forest_Gini_Tree250	Forest	0.8464	0.1536
	Forest_Gini (1)	Forest	0.8463	0.1537
	Forest_Entropy	Forest	0.8453	0.1547
	Forest_Base	Forest	0.8195	0.1805

The better performing method seems to be the model d-) according to the accuracy and F1 score which is gini split with a depth of 50.

## Visual Illustration



## Pipeline Comparison

Data Pipelines Pipeline Comparison Insights								
Filter		Data: Test						
<input type="checkbox"/>	Champion	Name	Algorithm Name	Pipeline Name	KS (Youden)	Sum of Frequencies	Area Under ROC	F1 Score
<input checked="" type="checkbox"/>		Forest_Gini_Depth50	Forest	Forest	0.669	11,939	0.920	0.780
<input type="checkbox"/>		Forest	Forest	Cluster	0.666	11,939	0.919	0.778
<input type="checkbox"/>		DT_Gini_C4.5_Depth15	Decision Tree	Dec_Tree	0.598	11,939	0.884	0.741
<input type="checkbox"/>		Ensemble	Ensemble	Bay_Net	0.567	11,939	0.869	0.725
<input type="checkbox"/>		Lo_Reg_Backward Impr...	Logistic Regression	Lo_Reg	0.561	11,939	0.864	0.716

When we compare the models, we can see that the Forest method with hyperparameters; split is Gini and the depth is 50 , gave the best model with an Area Under ROC 0.920 and with a F1 Score of 0.780

## Case Study

Tourism industry holds a big place in today's world. Millions of people go on vacations and make reservations from variety of hotels. This also means that, there is a big economic return in the tourism industry. From car vales to pool maintenance staff, from chefs to concierges there are a lot of expenditure needs to be made to hire the qualified people and this means money. The money comes from the customers who books or makes reservation to that hotel and to get the full price from the customers the reservation shouldn't be canceled. Cancellation of a reservation can lead to a money loss since there will be an empty room in the hotel and hotel managements make their plan for hiring people etc. according to how many people will stay at that hotel.



Using my best model which is forest method with Gini split and depth 50, we will be able to predict if a customer will cancel the reservation or not. This will allow hotel managements to book more than 1 customer to the same room and hotel managements should make up their own strategies if they want to book 5 customers to same room or 1 customer according to the model's prediction. This way, hotel management will reduce the chance of a room being empty. This means that that, hotel will not make loss in the revenue rather will increase revenue compare to the booking strategies used before implementing my model.

To be able to use my model, hotel management will need to hire a data scientist who has a deep knowledge of statistical learning a machine learning. Data scientist should be able to perform my model using various tools such as SAS Viya or coding through a coding language such as Python. Data scientist should also have to have a deep knowledge and understanding about these tools to apply my model in the best way. According to the past customer data that the hotel will provide, data scientist should be able to interpret it in the best way to include in to the model. Features that are shouldn't be included in the model, such as in our case reservation\_status and reservation\_status\_date, must be excluded to avoid building bad models.

The better the model we build, the better we will provide a satisfactory customer service but there will of course be a little chance that our model will make mistake since it is not %100 perfect. Decreasing this chance of mistake will directly dependent on the model we build and how much we can enhance it. My best model's area under ROC is 0.920 which is considered to be outstanding. I think applying my model to, hotel managements marketing and business strategy, will bring a huge benefit and will result in a increase in the revenue.

