ST 522/532 Final Project

Executive Summary

The original sample taken from a stratification by month results in a sample with a strong class imbalance. 98% of the data in the sample were for flights that were not canceled, and the remaining were for flights that were canceled. Using this sample resulted in models that only predict that a flight will not be canceled, resulting in a very good misclassification of only 2%. Therefore, if our only concern is predicting cancellations based on misclassification rates, then any model that always predicts no cancellation is a great model.

Another sample was taken, which consisted of all the cancellations and close to the same amount of non cancellations. The new sample was run with no prior distribution first, and the lowest misclassification rate was the optimal tree. (explain significant factors)

Then the models were run again with the prior probability, and the resulting misclassification rate was near 50%. This seems bad, but inspecting the resulting confusion table, shown in the appendix, will show that most of the models only predict that a flight will not be canceled. This was no different than the original sample.

The remaining models predict cancellations with low false positives. However, the optimal tree was found to be the best model based on the highest number of true positives and a low number of false positives. This model will outperform the previous models made using the original sample as well as the no prior probability models. This model was used to interpret the important factors in cancellation.

Introduction

This analysis is based on four data sets which are the following: flights, airlines, and airports. All of the data sets contain information about commercial air travel and are useful in our analysis in predicting flight cancellations. The airlines data set contains 14 observations and 2 variables, which provides each airline's name and airline's IATA code which is matched with the airline's airline code. The flight data set contains nearly 5.8 million observations and 31 variables that encompass a vast array of aspects of flights ranging from flight length to what day of the month the flight was scheduled. The hub dataset contains information on which airport each of the airlines have a hub at. Lastly, the airports dataset has 322 observations and 7 variables that describe the airports locations.

Methodology

Initially, the sample supplied by Professor Casselman to undergo data mining in SAS Enterprise was analyzed. Here, we noticed that there was a drastic split between canceled flights and non-cancelled flights. The prior probabilities showed that less than 2% of the observations were canceled. This discrepancy would lead to various inaccuracies in our models and we subsequently chose to resample from the original dataset. This resampling was performed in SAS Studio.

Firstly, the data was separated by cancellation and noncancellation. The noncancellation set was then sampled to match the amount that was in the cancellation set. A correlation table was then produced to see if there was any multicollinearity among the predictor variables. We found out that variables scheduled_time and distance are highly correlated with a correlation coefficient of .98. A frequency table was also produced to ensure the amount of canceled and non canceled flights were equal. Additionally, a two way frequency table was created to check the missing values of delay variables. The results showed that departure delay values were missing for 95% of all canceled flights and 0% for flights not canceled. Departure delay values are a summation of all the delays before the flight leaves, so a missing value for departure delay means no delay values were recorded. Since 95% of the canceled flights did not have value, we could not impute and therefore rejected departure delay and consequently all delays. A summary table of airports was also created through a proc sql statement grouping flights by airport. Numerous columns were created describing statistics and 95% confidence intervals of the statistics. The columns created were percent diverted, percent canceled, average departure time, and average arrival time with their respective lower and upper bounds of a 95% confidence interval. Average departure time and Average arrival time are relative to their scheduled times. A summary table for airlines was also created in a similar fashion grouping by airline names. The columns created were the same as the airport summary except no confidence intervals were created. The amount of flights per airline provides a very good estimator of the statistics created. The airport summary table was then used to make another table that consisted of airports that had above average percent canceled. This table was made by using a subquery of average percent canceled. A table of monthly cancellations was made with the number of cancellation and percent of cancellation. This table shows that February had the most cancellations along with the highest percentage of cancellations. Moreover, a table of day of week was made with total flights and percent canceled. The table shows us that Thursday has the highest number of flights and Sunday has the highest percentage canceled at 2%. A table of hub and percent cancellation was built to determine if cancellation happens more when the flight's airline has a hub at the location. Using this table, we see that flights without hubs have slightly more cancellations than those that are at a hub.

Quickly on we noticed that the delay variables are almost all missing 98% of the time for when the flight is canceled. With this we concluded that delay associated variables should be rejected in the model to predict cancellations. Another issue we found was that distance and scheduled time were correlated so we chose to reject scheduled time. After the data split we started the model training.

Neural Networks

We ran the data through four neural networks with the default 3 hidden units, 5,10, and 25 hidden units. We also wanted to run an ensemble neural network so all the individual neural networks are connected to a control point named "Control Point NN". We made ensemble models with all of the hidden unit models, another with 5-25 hidden units and another with the 10 and 25 hidden units. All of the individual neural networks as well as the ensemble models are connected to the model comparison node.

Tree-Based Models

We ran an optimal decision tree and connected the tree to the model comparison. We also ran a random forest so we could get the OOB statistics and have that comparison in the model comparison. With the random forest, we also did a bagging and boosting model. The boosting model was done using the gradient boosting node with an assessment measure of misclassification. The bagging model was done using the start/end groups nodes to make the samples. The samples were run through an optimal tree and then sent to the end groups node. We then connected the various end sample models to the model comparison node as well as an ensemble node named “Bagging tree ensemble." This bagging ensemble was then connected to the model comparison as well.

Regression Models

Since we don’t have a lot of interval inputs, there was a concern about possible interactions in the model. To test this, we chose to run various regression models using different model selection techniques for models with and without interactions. To simplify the arrows in the flow chart, we connected the data partition node to a control point. This was then connected to a total of six logistic regression models. Three models used stepwise, forward, and backwards selection without interaction terms included, and the other three used stepwise, forward, and backwards selection with interaction terms included. We then ran two regression ensemble models, combining those with interaction into the “reg BIC yes int” node and those without interaction into the “reg BIC no int” node. All six individual regression models were connected to a control point named “Control Point reg models” and then connected to the model comparison. Both regression ensemble models were also connected to the model comparison.

Different Sample

As mentioned earlier, there was a large difference between our primary and secondary targets. We took another sample to attempt to remedy this issue in the model building. Hub and quarter were added to the list of potential predictors. We then ran the same models as mentioned with the new data and compared the results.

Results

A breakdown of the various models run for model comparison for each of the data samples can be seen in the appendix. As mentioned earlier, the optimal decision tree with the re-sampled 50/50 split and prior probabilities was our best model. The new data sample, with the addition of the prior probabilities, allowed the model to be more accurately trained to predict canceled flights instead of non-canceled flights. The final and best model had a validation misclassification rate of 0.484459. It also has the highest count of true positives, with a validation true positive count of 1130. Additionally, it had the most false positives out of the models with 34 false positives in the validation set. No other models really over-predicted the canceled flights as much as the optimal tree, however this hesitance in the other models means that it was still over-predicting non-cancelled flights. When the optimal tree was examined further, it was seen that there were two main factors in the tree that determined a prediction of cancellation with decent accuracy. The first being if the flight is in the month of February, the day of the month is 23 or greater with certain airlines, and if it's either a Friday or Sunday. The second factor, while still important, is still less practical than the first factor. The second factor is if it is the fifth day of March and a certain airline.

The second best model was the logistic regression model that utilized Bayesian information criterion with backwards variable selection. Despite this, the model is quite uninterpretable. The amount of nominal variables and interactions between them can not be interpreted with significant meaning. With this, we again focus on the optimal tree for the conclusion.

Discussion/Conclusion

Based on the sample that was roughly 50% canceled and 50% not canceled observations, the best misclassification rate was 22%. This shows that available predictors do not predict cancellation that well. This would also explain why the models run using the original sample that had less than 2% of the flights canceled were not predicting any cancellations.

This is probably due to quality control measures used by airlines to reduce cancellations. If an airline finds a significant factor that increases cancellations, they would try to fix it as cancellations are extremely costly for the airlines. This is also shown in the data of the delays. Over 98% of the observations that were canceled had missing delay variables and those that were delayed were marked as non-canceled. This shows that if a flight had to be delayed, it very rarely led to a cancellation. If the flight was delayed past the point of the airport closing times, it was considered to be a National Air System delay, not canceled. While major delays mean the passengers are entitled to some compensation, most delays, if any, were not more costly than refunding an entire flight.

A possible further analysis could be done on the relationship between flight cancellations and delays and the type of plane used on the flights. It would be interesting to see if the boeings fly better in certain months compared to other types. It would also be interesting to include average flight ticket prices into this analysis. Do cheaper flights get delayed or canceled more often than the more expensive flights? With more time and resources, much more can be done with analysis.

Appendix

Original sample

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	Neur…	Neur…	Neural Network 10h	CA…		0.01609
	Neur…	Neur…	Neural Network 5h	CA…		0.016123
	HPD…	HPD…	HP Forest	CA…		0.016132
	Ens…	Ens…	10&25	CA…		0.016132
	Ens…	Ens…	5-25	CA…		0.016132
	Ens…	Ens…	Ensemble all hh	CA…		0.016132
	Reg12	Reg12	regbackBICyesint	CA…		0.016132
	Ens…	Ens…	reg BIC yes int	CA…		0.016132
	Reg11	Reg11	regstepBICyesint	CA…		0.016132
	Reg13	Reg13	regforwBICyesint	CA…		0.016132
	Neur…	Neur…	Neural Network 25h	CA…		0.016132
	Reg10	Reg10	regforwBICnoint	CA…		0.016132
	Reg8	Reg8	regstepBICnoint	CA…		0.016132
	Reg9	Reg9	regbackBICnoint	CA…		0.016132
	Ens…	Ens…	reg BIC no int	CA…		0.016132
	Neural	Neural	Neural Network	CA…		0.016132
	Boost	Boost	Gradient Boosting	CA…		0.016132
	Tree	Tree	optimal tree	CA…		0.016132
	Ens…	Ens…	Bagging tree ensem…	CA…		0.016132

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree	optimal tree	TRAIN	CANCELLED		2887	177101	0	0
Tree	optimal tree	VALIDATE	CANCELLED		1936	118076	0	0
Boost	Gradient Boosting	TRAIN	CANCELLED		2887	177101	0	0
Boost	Gradient Boosting	VALIDATE	CANCELLED		1936	118076	0	0
HPDMForest	HP Forest	TRAIN	CANCELLED		2887	177101	0	0
HPDMForest	HP Forest	VALIDATE	CANCELLED		1936	118076	0	0
Ensmbl3	5-25	TRAIN	CANCELLED		2887	177101	0	0
Ensmbl3	5-25	VALIDATE	CANCELLED		1936	118076	0	0
Ensmbl	Ensemble all hh	TRAIN	CANCELLED		2887	177101	0	0
Ensmbl	Ensemble all hh	VALIDATE	CANCELLED		1936	118076	0	0
Neural3	Neural Network 25h	TRAIN	CANCELLED		2887	177101	0	0
Neural3	Neural Network 25h	VALIDATE	CANCELLED		1936	118076	0	0
Neural2	Neural Network 10h	TRAIN	CANCELLED		2884	177093	8	3
Neural2	Neural Network 10h	VALIDATE	CANCELLED		1928	118073	3	8
Neural	Neural Network	TRAIN	CANCELLED		2887	177101	0	0
Neural	Neural Network	VALIDATE	CANCELLED		1936	118076	0	0
Neural4	Neural Network 5h	TRAIN	CANCELLED		2884	177100	1	3
Neural4	Neural Network 5h	VALIDATE	CANCELLED		1934	118075	1	2
Ensmbl4	10&25	TRAIN	CANCELLED		2887	177101	0	0
Ensmbl4	10&25	VALIDATE	CANCELLED		1936	118076	0	0
Ensmbl2	Bagging tree ensemble	TRAIN	CANCELLED		2887	177101	0	0
Ensmbl2	Bagging tree ensemble	VALIDATE	CANCELLED		1936	118076	0	0
Ensmbl8	reg BIC yes int	TRAIN	CANCELLED		2887	177101	0	0
Ensmbl8	reg BIC yes int	VALIDATE	CANCELLED		1936	118076	0	0
Reg8	regstepBICnoint	TRAIN	CANCELLED		2887	177101	0	0
Reg8	regstepBICnoint	VALIDATE	CANCELLED		1936	118076	0	0
Reg9	regbackBICnoint	TRAIN	CANCELLED		2887	177101	0	0
Reg9	regbackBICnoint	VALIDATE	CANCELLED		1936	118076	0	0
Reg10	regforwBICnoint	TRAIN	CANCELLED		2887	177101	0	0
Reg10	regforwBICnoint	VALIDATE	CANCELLED		1936	118076	0	0
Reg11	regstepBICyesint	TRAIN	CANCELLED		2887	177101	0	0
Reg11	regstepBICyesint	VALIDATE	CANCELLED		1936	118076	0	0
Reg12	regbackBICyesint	TRAIN	CANCELLED		2887	177101	0	0
Reg12	regbackBICyesint	VALIDATE	CANCELLED		1936	118076	0	0
Reg13	regforwBICyesint	TRAIN	CANCELLED		2887	177101	0	0
Reg13	regforwBICyesint	VALIDATE	CANCELLED		1936	118076	0	0
Ensmbl7	reg BIC no int	TRAIN	CANCELLED		2887	177101	0	0
Ensmbl7	reg BIC no int	VALIDATE	CANCELLED		1936	118076	0	0

50/50 no prior

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	HPD...	HPD...	HP Forest	CAN...		0.22937
	Ens...	Ens...	5-25	CAN...		0.257757
	Ens...	Ens...	10&25	CAN...		0.259049
	Ens...	Ens...	Ensemble all hh	CAN...		0.260369
	Neur...	Neur...	Neural Network 10h	CAN...		0.26305
	Neur...	Neur...	Neural Network 25h	CAN...		0.263273
	Neur...	Neur...	Neural Network 5h	CAN...		0.268483
	Ens...	Ens...	reg BIC yes int	CAN...		0.285587
	Reg12	Reg12	regbackBICyesint	CAN...		0.285726
	Reg11	Reg11	regstepBICyesint	CAN...		0.285823
	Reg13	Reg13	regforwBICyesint	CAN...		0.285823
	Neural	Neural	Neural Network	CAN...		0.288839
	Ens...	Ens...	Bagging tree ensemble	CAN...		0.294174
	Tree	Tree	optimal tree	CAN...		0.307596
	Ens...	Ens...	reg BIC no int	CAN...		0.318809
	Reg10	Reg10	regforwBICnoint	CAN...		0.318809
	Reg8	Reg8	regstepBICnoint	CAN...		0.318809
	Reg9	Reg9	regbackBICnoint	CAN...		0.318809
	Boost	Boost	Gradient Boosting	CAN...		0.499687

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree	optimal tree	TRAIN	CANCELLED		12776	33599	20393	41145
Tree	optimal tree	VALIDATE	CANCELLED		8597	22467	13541	27366
Boost	Gradient Boosting	TRAIN	CANCELLED		53921	53992	0	0
Boost	Gradient Boosting	VALIDATE	CANCELLED		35963	36008	0	0
HPDMForest	HP Forest	TRAIN	CANCELLED		12488	43002	10990	41433
HPDMForest	HP Forest	VALIDATE	CANCELLED		8706	28170	7838	27257
Ensmbl3	5-25	TRAIN	CANCELLED		14572	40779	13213	39349
Ensmbl3	5-25	VALIDATE	CANCELLED		9776	27233	8775	26187
Ensmbl	Ensemble all hh	TRAIN	CANCELLED		14442	40414	13578	39479
Ensmbl	Ensemble all hh	VALIDATE	CANCELLED		9729	26998	9010	26234
Neural3	Neural Network 25h	TRAIN	CANCELLED		13946	39580	14412	39975
Neural3	Neural Network 25h	VALIDATE	CANCELLED		9326	26386	9622	26637
Neural2	Neural Network 10h	TRAIN	CANCELLED		14472	39973	14019	39449
Neural2	Neural Network 10h	VALIDATE	CANCELLED		9713	26789	9219	26250
Neural	Neural Network	TRAIN	CANCELLED		14896	37890	16102	39025
Neural	Neural Network	VALIDATE	CANCELLED		10020	25240	10768	25943
Neural4	Neural Network 5h	TRAIN	CANCELLED		15401	40348	13644	38520
Neural4	Neural Network 5h	VALIDATE	CANCELLED		10283	26968	9040	25680
Ensmbl4	10&25	TRAIN	CANCELLED		14152	40314	13678	39769
Ensmbl4	10&25	VALIDATE	CANCELLED		9478	26842	9166	26485
Ensmbl2	Bagging tree ensemble	TRAIN	CANCELLED		16431	38851	15141	37490
Ensmbl2	Bagging tree ensemble	VALIDATE	CANCELLED		11124	25960	10048	24839
Ensmbl8	reg BIC yes int	TRAIN	CANCELLED		15161	38369	15623	38760
Ensmbl8	reg BIC yes int	VALIDATE	CANCELLED		10172	25626	10382	25791
Reg8	regstepBICnoint	TRAIN	CANCELLED		17159	36649	17343	36762
Reg8	regstepBICnoint	VALIDATE	CANCELLED		11378	24441	11567	24585
Reg9	regbackBICnoint	TRAIN	CANCELLED		17159	36649	17343	36762
Reg9	regbackBICnoint	VALIDATE	CANCELLED		11378	24441	11567	24585
Reg10	regforwBICnoint	TRAIN	CANCELLED		17159	36649	17343	36762
Reg10	regforwBICnoint	VALIDATE	CANCELLED		11378	24441	11567	24585
Reg11	regstepBICyesint	TRAIN	CANCELLED		15158	38385	15607	38763
Reg11	regstepBICyesint	VALIDATE	CANCELLED		10178	25615	10393	25785
Reg12	regbackBICyesint	TRAIN	CANCELLED		15116	38345	15647	38805
Reg12	regbackBICyesint	VALIDATE	CANCELLED		10159	25603	10405	25804
Reg13	regforwBICyesint	TRAIN	CANCELLED		15158	38385	15607	38763
Reg13	regforwBICyesint	VALIDATE	CANCELLED		10178	25615	10393	25785
Ensmbl7	reg BIC no int	TRAIN	CANCELLED		17159	36649	17343	36762
Ensmbl7	reg BIC no int	VALIDATE	CANCELLED		11378	24441	11567	24585

50/50 with prior

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	Tree	Tree	optimal tree	CAN...		0.484459
	Neur...	Neur...	Neural Network 10h	CAN...		0.497034
	Neur...	Neur...	Neural Network 25h	CAN...		0.498854
	Neur...	Neur...	Neural Network 5h	CAN...		0.498881
	Reg12	Reg12	regbackBICyesint	CAN...		0.499479
	Ens...	Ens...	10&25	CAN...		0.499479
	Ens...	Ens...	5-25	CAN...		0.499562
	Ens...	Ens...	reg BIC yes int	CAN...		0.499562
	Reg11	Reg11	regstepBICyesint	CAN...		0.499618
	Reg13	Reg13	regforwBICyesint	CAN...		0.499618
	HPD...	HPD...	HP Forest	CAN...		0.499687
	Neural	Neural	Neural Network	CAN...		0.499687
	Reg10	Reg10	regforwBICnoint	CAN...		0.499687
	Reg8	Reg8	regstepBICnoint	CAN...		0.499687
	Reg9	Reg9	regbackBICnoint	CAN...		0.499687
	Boost	Boost	Gradient Boosting	CAN...		0.499687
	Ens...	Ens...	Ensemble all hh	CAN...		0.499687
	Ens...	Ens...	reg BIC no int	CAN...		0.499687
	Ens...	Ens...	Bagging tree ensem...	CAN...		0.499687

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree	optimal tree	TRAIN	CANCELLED		52180	53961	31	1741
Tree	optimal tree	VALIDATE	CANCELLED		34833	35974	34	1130
Boost	Gradient Boosting	TRAIN	CANCELLED		53921	53992	0	0
Boost	Gradient Boosting	VALIDATE	CANCELLED		35963	36008	0	0
HPDMForest	HP Forest	TRAIN	CANCELLED		53921	53992	0	0
HPDMForest	HP Forest	VALIDATE	CANCELLED		35963	36008	0	0
Ensmbl3	5-25	TRAIN	CANCELLED		53897	53991	1	24
Ensmbl3	5-25	VALIDATE	CANCELLED		35954	36008	.	9
Ensmbl	Ensemble all hh	TRAIN	CANCELLED		53921	53992	0	0
Ensmbl	Ensemble all hh	VALIDATE	CANCELLED		35963	36008	0	0
Neural3	Neural Network 25h	TRAIN	CANCELLED		53828	53988	4	93
Neural3	Neural Network 25h	VALIDATE	CANCELLED		35903	36008	.	60
Neural2	Neural Network 10h	TRAIN	CANCELLED		53658	53990	2	263
Neural2	Neural Network 10h	VALIDATE	CANCELLED		35768	36004	4	195
Neural	Neural Network	TRAIN	CANCELLED		53921	53992	0	0
Neural	Neural Network	VALIDATE	CANCELLED		35963	36008	0	0
Neural4	Neural Network 5h	TRAIN	CANCELLED		53832	53990	2	89
Neural4	Neural Network 5h	VALIDATE	CANCELLED		35905	36008	.	58
Ensmbl4	10&25	TRAIN	CANCELLED		53897	53992	0	24
Ensmbl4	10&25	VALIDATE	CANCELLED		35948	36008	0	15
Ensmbl2	Bagging tree ensemble	TRAIN	CANCELLED		53921	53992	0	0
Ensmbl2	Bagging tree ensemble	VALIDATE	CANCELLED		35963	36008	0	0
Ensmbl8	reg BIC yes int	TRAIN	CANCELLED		53911	53992	0	10
Ensmbl8	reg BIC yes int	VALIDATE	CANCELLED		35954	36008	0	9
Reg8	regstepBICnoint	TRAIN	CANCELLED		53921	53992	0	0
Reg8	regstepBICnoint	VALIDATE	CANCELLED		35963	36008	0	0
Reg9	regbackBICnoint	TRAIN	CANCELLED		53921	53992	0	0
Reg9	regbackBICnoint	VALIDATE	CANCELLED		35963	36008	0	0
Reg10	regforwBICnoint	TRAIN	CANCELLED		53921	53992	0	0
Reg10	regforwBICnoint	VALIDATE	CANCELLED		35963	36008	0	0
Reg11	regstepBICyesint	TRAIN	CANCELLED		53915	53992	0	6
Reg11	regstepBICyesint	VALIDATE	CANCELLED		35958	36008	0	5
Reg12	regbackBICyesint	TRAIN	CANCELLED		53902	53992	0	19
Reg12	regbackBICyesint	VALIDATE	CANCELLED		35948	36008	0	15
Reg13	regforwBICyesint	TRAIN	CANCELLED		53915	53992	0	6
Reg13	regforwBICyesint	VALIDATE	CANCELLED		35958	36008	0	5
Ensmbl7	reg BIC no int	TRAIN	CANCELLED		53921	53992	0	0
Ensmbl7	reg BIC no int	VALIDATE	CANCELLED		35963	36008	0	0

```
proc corr data=proj.flights;
	var scheduled_time distance;
 run;
```

The CORR Procedure

2 Variables:	SCHEDULED_TIME DISTANCE

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
SCHEDULED_TIME	5819073	141.68589	75.21058	824480546	18.00000	718.00000
DISTANCE	5819079	822.35649	607.78429	4785357409	21.00000	4983

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	SCHEDULED_TIME	**DISTANCE**
SCHEDULED_TIME	1.00000 5819073	0.98434 <.0001 5819073
DISTANCE	0.98434 <.0001 5819073	1.00000 5819079

```
proc freq data=proj.flights;
	table cancelled;
```

```
Run;
```

The FREQ Procedure

CANCELLED	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5729195	98.46	5729195	98.46
1	89884	1.54	5819079	100.00

```
proc sql;
	create table work.flightsairports as
	select *from proj.flights left join proj.airports
	on flights.origin_airport = airports.iata_code;
quit;

proc sql;
	create table work.flights as
	select *from work.flightsairports as a left join proj.airlines
	on a.airline = airlines.iata_code;
quit;

proc sql;
```

```
create table airline_stats as
    select air_line,
    count(*) as number_of_flights,
    sum(diverted)/count(*) as proportion_diverted,
    sum(cancelled)/count(*) as proportion_cancelled,
    sum(departure_delay)/count(*) as avg_departure_time,
    sum(arrival_delay)/count(*) as avg_arrival_time
    from work.flights
    group by air_line
    order by proportion_cancelled desc;
quit;
```

Airline summary table

Total rows: 14 Total columns: 6 Rows 1-14

	air_line	number_of_flights	proportion_div
1	American Eagle Airlines Inc.	294632	0.002769
2	Atlantic Southeast Airlines	571977	0.003486
3	US Airways Inc.	198715	0.002138
4	Spirit Air Lines	117379	0.001550
5	Skywest Airlines Inc.	588353	0.00268
6	JetBlue Airways	267048	0.00273
7	American Airlines Inc.	725984	0.00293
8	United Air Lines Inc.	515723	0.002691
9	Southwest Airlines Co.	1261855	0.002701
10	Virgin America	61903	0.00195
11	Frontier Airlines Inc.	90836	0.001739
12	Delta Air Lines Inc.	875881	0.00203
13	Alaska Airlines Inc.	172521	0.002393
14	Hawaiian Airlines Inc.	76272	0.000786

```
proc sql;

create table origin_airport_stats as

        select airport,

        count(*) as number_of_flights,

        sum(diverted)/count(*) as proportion_diverted,
```

sum(diverted)/count(*)-
1.96*(sqrt(((sum(diverted)/count(*))*(1-sum(diverted)/count(*)))/count(*))) as
lower_bound_diverted,

sum(diverted)/count(*)+
1.96*(sqrt(((sum(diverted)/count(*))*(1-sum(diverted)/count(*)))/count(*))) as
higher_bound_diverted,

sum(cancelled)/count(*) as proportion_cancelled,

sum(cancelled)/count(*)-
1.96*(sqrt(((sum(cancelled)/count(*))*(1-sum(cancelled)/count(*)))/count(*))) as
lower_bound_cancelled,

sum(cancelled)/count(*)+
1.96*(sqrt(((sum(cancelled)/count(*))*(1-sum(cancelled)/count(*)))/count(*))) as
higher_bound_cancelled,

sum(departure_delay)/count(*) as avg_departure_time,

sum(departure_delay)/count(*)-1.96*(std(departure_delay)/count(*)) as
lower_bound_departure,

sum(departure_delay)/count(*)+1.96*(std(departure_delay)/count(*)) as
higher_bound_departure,

sum(arrival_delay)/count(*) as avg_arrival_time,

sum(arrival_delay)/count(*)-1.96*(std(arrival_delay)/count(*)) as lower_bound_arrival,

sum(arrival_delay)/count(*)+1.96*(std(arrival_delay)/count(*)) as higher_bound_arrival

from proj.complete

group by airport

order by proportion_cancelled desc;

quit;

Airport summary table

Total rows: 323 Total columns: 14 Rows 1-100

	number_of_flights	proportion_diverted	lower_bound_diverted
ıal Airport	34	0	0
·port	156	0	0
port	956	0.0010460251	-0.001003112
rport	525	0.0247619048	0.0114688813
›rtÂ (Jack McNamara Fi	190	0.0052631579	-0.005025449
rport	3562	0.0064570466	0.0038266611
ort	667	0	0
	96	0.0104166667	-0.009893385
irport	812	0.0123152709	0.0047293343
›nal AirportÂ (St. Au	155	0	0
al Airport	1331	0.0007513148	-0.000720709
	962	0	0
ıorial Airport	668	0.005988024	0.0001373565
rt	666	0	0
irportÂ (Southeast Te	973	0.0051387461	0.0006460247

proc sql;

select airport, percent_cancelled from origin_airport_stat

where percent_cancelled > (select avg(percent_cancelled) from origin_airport_stats)

order by percent_cancelled desc;

Quit;

Airport subquery above average proportion cancelled

AIRPORT	percent_cancelled
Ithaca Tompkins Regional Airport	0.117647
Mammoth Yosemite Airport	0.102564
Friedman Memorial Airport	0.09205
Devils Lake Regional Airport	0.087619
Del Norte County AirportÂ (Jack McNamara Fi	0.084211
Aspen-Pitkin County Airport	0.077485
Muskegon County Airport	0.073463
Adak Airport	0.072917
Jamestown Regional Airport	0.07266
Northeast Florida Regional AirportÂ (St. Au	0.070968
Lawton-Fort Sill Regional Airport	0.070624

```
proc sql;
	select month, sum(cancelled) as cancellations , sum(cancelled)/count(*) as percent_cancelled
	from proj.complete
	group by month
	order by cancellations desc;
```

Quit;

Month summary table

MONTH	cancellations	percent_cancelled
2	20517	0.047804
1	11982	0.025495
3	11002	0.021816
6	9120	0.018099
12	8063	0.016825
5	5694	0.011457
8	5052	0.009895
7	4806	0.00923
11	4599	0.009828
4	4520	0.009317
10	2454	0.005048
9	2075	0.004463

```
data work.flights;

    set proj.flights;

    if departure_delay=. then  delay=0 ;

    else if departure_delay >=0  then delay=1;
```

```
        else  delay=-1;

run;



proc freq data=work.flights;

        tables cancelled*delay;

Run;
```

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of CANCELLED by delay

CANCELLED	delay -1	delay 0	delay 1	Total
0	3276871 56.31 57.20 99.97	0 0.00 0.00 0.00	2452324 42.14 42.80 99.89	5729195 98.46
1	1077 0.02 1.20 0.03	86153 1.48 95.85 100.00	2654 0.05 2.95 0.11	89884 1.54
Total	3277948 56.33	86153 1.48	2454978 42.19	5819079 100.00

(where 0=missing value)

```
data work.quarters;
	set proj.flightairline;
	if month in (1,2,3) then quarter=1;
	else if month in (4,5,6) then quarter=2;
	else if month in (7,8,9) then quarter=3;
	else quarter=4;
run;

data work.cancelled work.notcancelled;
	set work.quarters;
	if cancelled=1 then output work.cancelled;
	else output work.notcancelled;
run;

proc surveyselect data=work.notcancelled  samprate==.6365 out=work.notcancelledsample;
run;

data proj.newsamp;
	set work.cancelled work.notcancelledsample;
run;
```

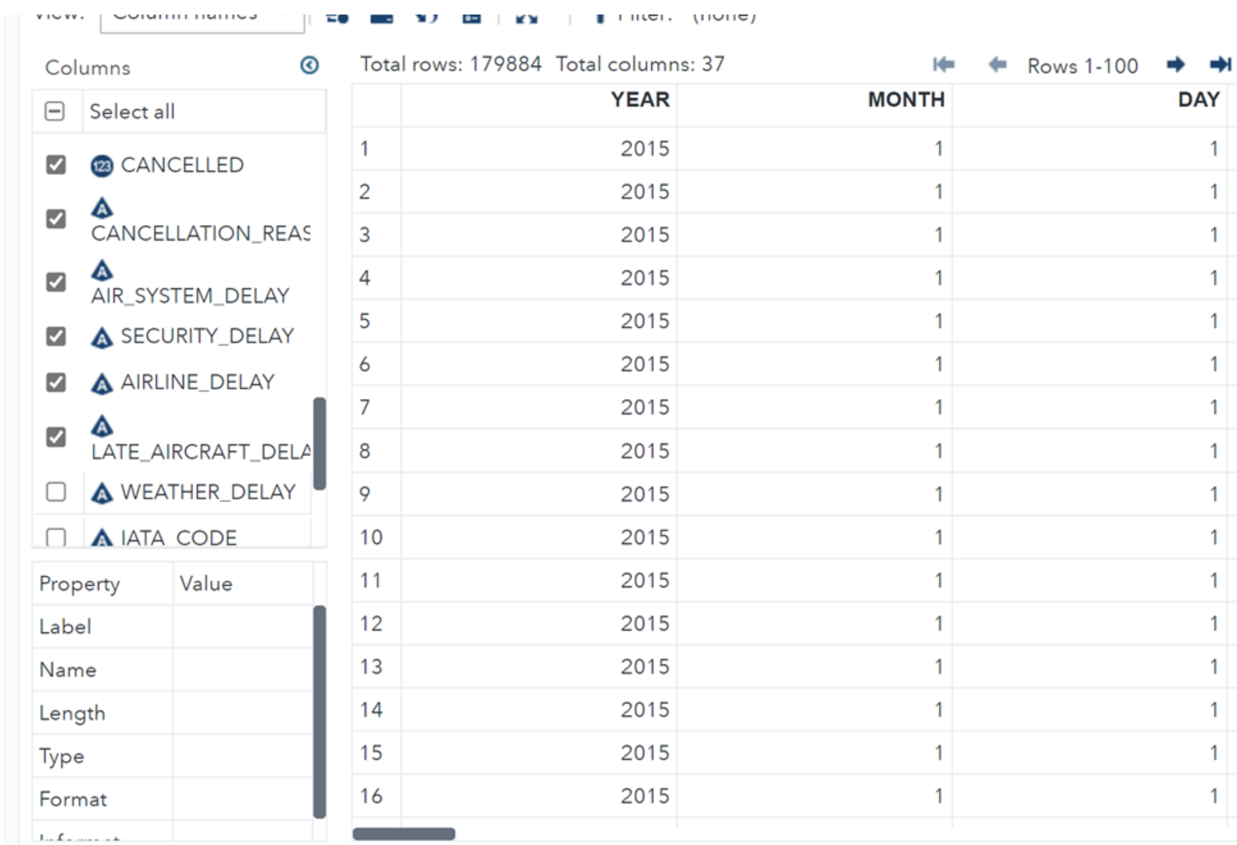
```
data a; set proj.hubs;
  do i = 1 to 12;
        apt_code = scan(hub_airports, i, " ");
        output;
  end;
run;

proc sql;
        create table newsamp1 as
        select * from proj.newsamp left join  work.a
        on newsamp.airline = a.airline And
        newsamp.origin_airport = a.apt_code
        ;
quit;

proc sort data=work.newsamp1;
        by year month day;
run;
```

```
data proj.newsamp1(drop= i  samplingweight selectionprob );
	set work.newsamp1;
	if apt_code= " " then hub=0;
	else hub=1;
Run;
```

Creating a sample



	YEAR	MONTH	DAY
1	2015	1	1
2	2015	1	1
3	2015	1	1
4	2015	1	1
5	2015	1	1
6	2015	1	1
7	2015	1	1
8	2015	1	1
9	2015	1	1
10	2015	1	1
11	2015	1	1
12	2015	1	1
13	2015	1	1
14	2015	1	1
15	2015	1	1
16	2015	1	1

```
proc sql;
	select hub, sum(cancelled)/count(*) as proportion_cancelled
	from proj.complete
```

```
	group by hub;
Quit;
```

Hub summary table

hub	proportion_cancelled
0	0.016026
1	0.014774