**Final Project**

The data for this project comes from the mobile advertising space. In order to encourage consumers to install its app (e.g. a game), an app developer advertises its app on other apps (e.g., other games) through a mobile advertising platform. Consumers viewing these ads on these other apps can click on the ad to install the app from the developer. We will refer to the advertising app developer as the advertiser. See figure below.



Consumer

Install

Publisher 2

Advertiser

Advertising

Platform

.

.

.

Consumer

.

.

.

Consumer

Not Install

Publisher k-1

Consumer

Publisher k

Publisher 1

The dataset for this project contains data about ads from one particular advertiser through multiple publishers. Each observation corresponds to one ad shown to a consumer on a particular publisher app. The observation contains information about the publisher id, consumer’s device characteristics, and whether the advertiser’s app was installed or not. The description of the variables are given below.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| publisher\_id\_class | Categorical | Publisher Id |
| device\_make\_class | Categorical | Device Manufacturer |
| device\_platform\_class | Categorical | Phone OS Type (iPhone / Android) |
| device\_os\_class | Categorical | Phone OS Version |
| device\_height | Numerical | Display Height (in pixels) |
| device\_width | Numerical | Display Width (in pixels) |
| Resolution | Numerical | Display Resolution (pixels per inch) |
| device\_volume | Numerical | Device Volume when Ad was displayed |
| Wifi | Numerical | Whether WiFi was enabled when ad was displayed (Yes = 1, No  = 0) |
| Install | Binary | Whether Consumer Installed Advertiser’s App (Yes = 1, No = 0) |

**DATA PREPARATION:**

The device platform variable was converted to a binary 0,1 variable. 1=’IOS’

Publisher\_ID and Device\_make were converted to class variables

# Part I. Logistic Regression Analysis

1. The advertiser needs to determine how much to pay for placing ad depending on the publisher and on the consumer characteristics. The optimal payment is proportional to the probability that a consumer seeing the ad will install the ad.

Develop a logistic regression model to estimate the probability of installing the ad based on publisher and consumer characteristics. Present only the final model and explain the procedure and different measures you have used to come up with this model.

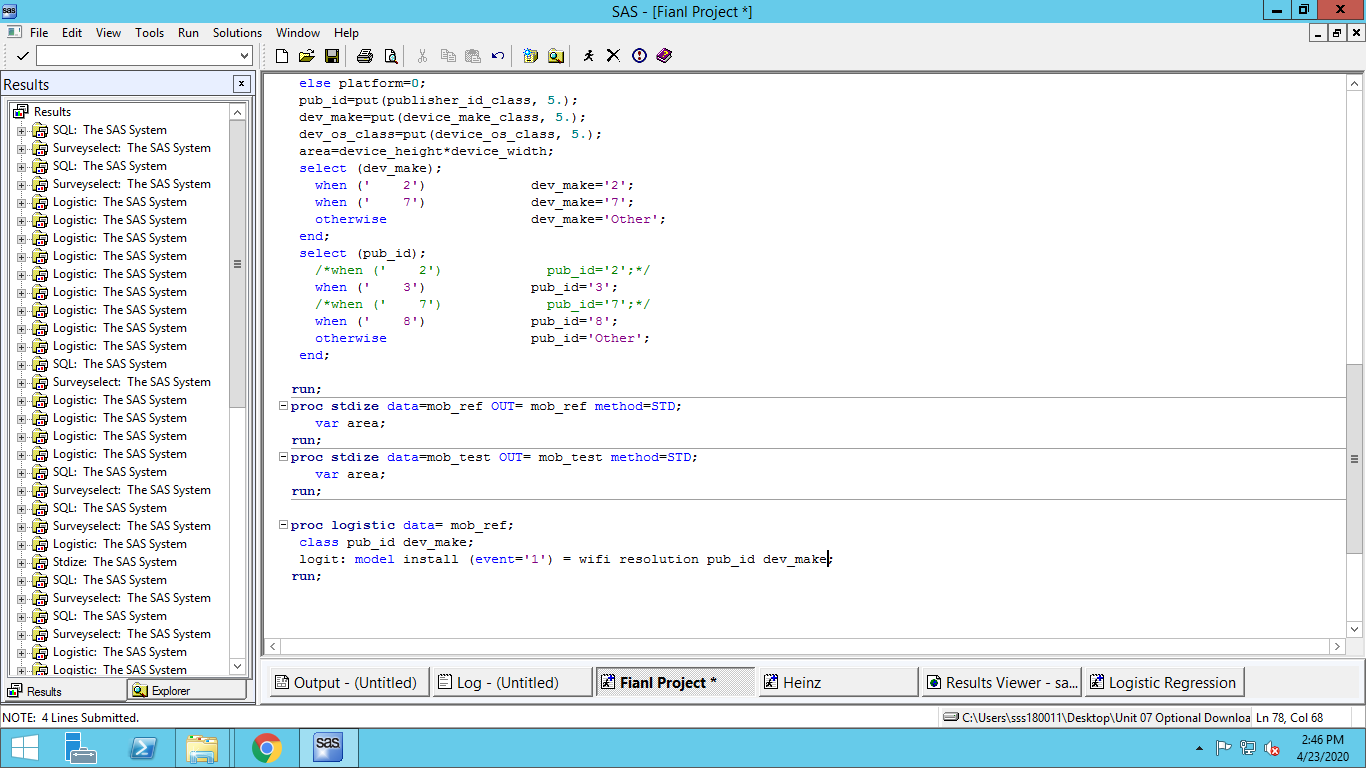
**Even though the percentage of positives/negatives is low, since there are about 1008 positive observations usual MLE estimation was used and no modification done to dataset.**

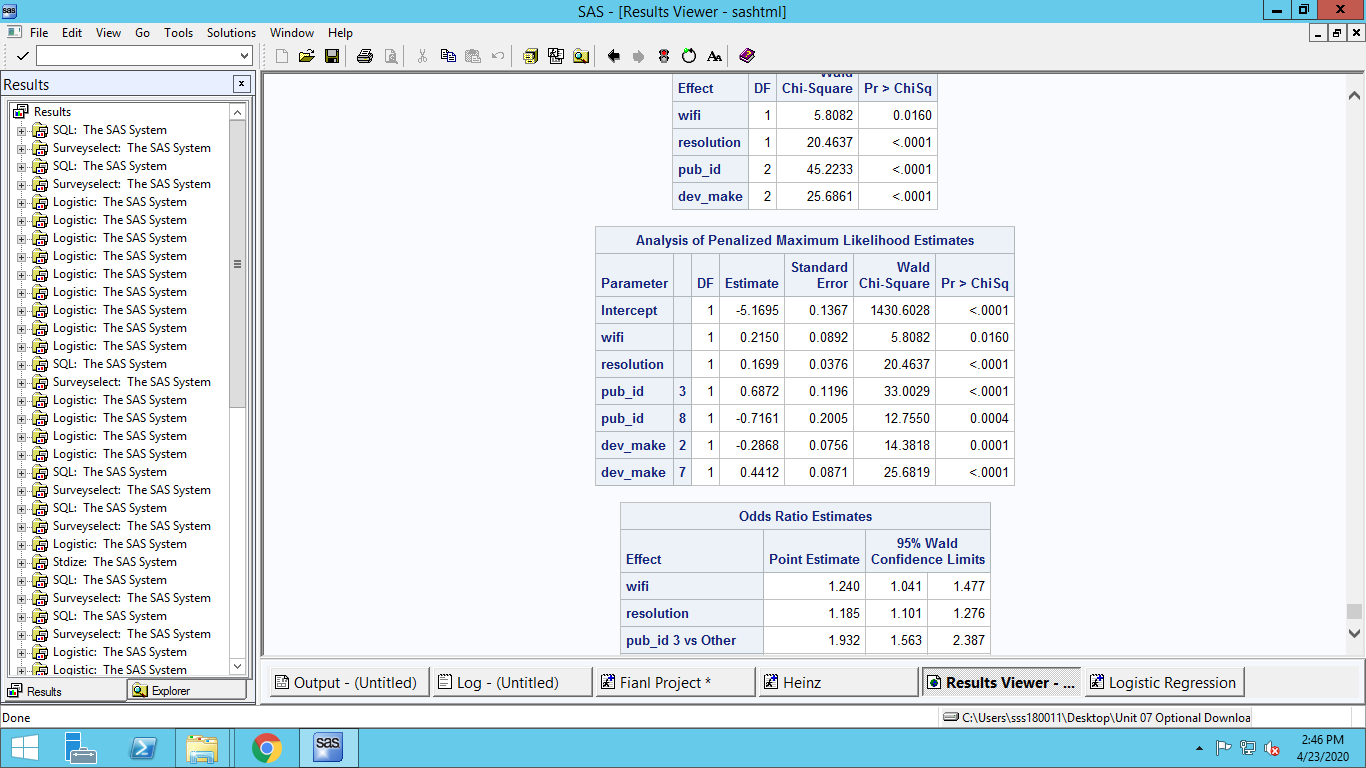
**Device Volume, platform and device OS\_class are not statistically significant, hence they were removed.**

**Only the publisher\_ID of 3,8 and device\_make of 2,7 were statistically significant, hence only they were kept.**

**Resolution and device\_width were also not statistically significant, hence were removed. The parameter value of device\_height was too low to be of any practical value, hence it was removed. An area variable created by multiplying height and width. The area variable was statistically signifivant.**

**When logistic model was created area variable was set to zero since there was some collinearity between area and resolution variable. Hence area was removed**

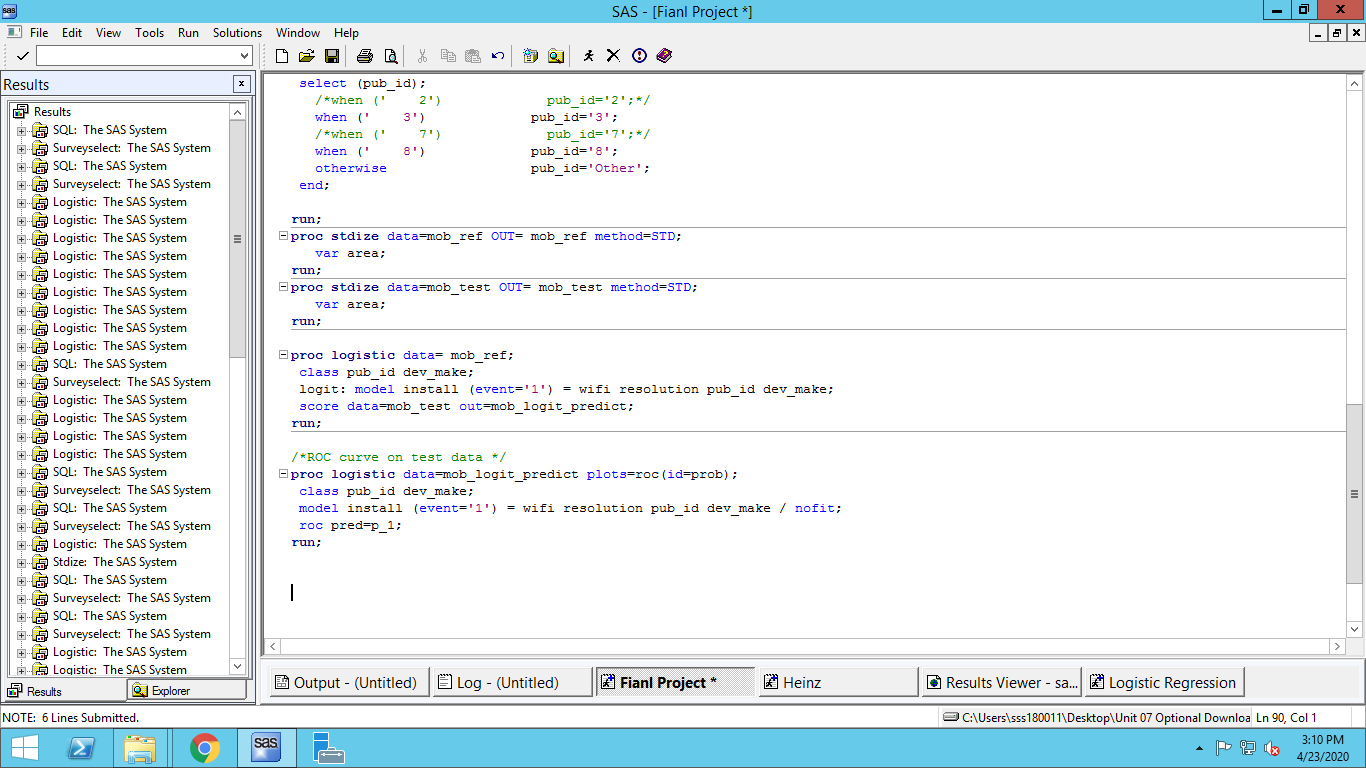


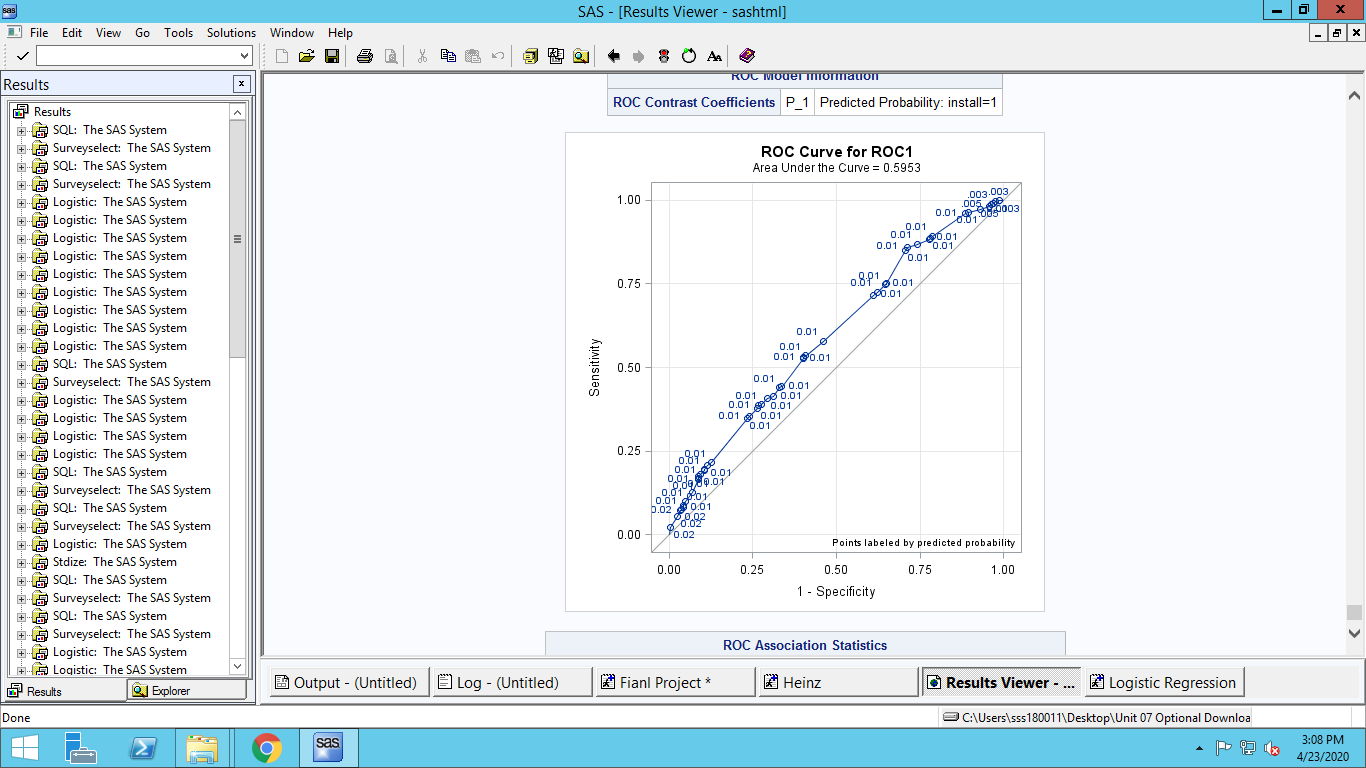


**THE MODEL: INSTALL = -3.6433\*Intercept + 0.3183\*WIFI + 0.1527\*Resolution + 0.3072\*Pub\_ID\_2 + 0.7137\*Pub\_ID\_3 + -0.4759\*Pub\_ID\_7 + -0.6564\* Pub\_ID\_8 + -0.2541\* dev\_make\_2 + 0.4217\* dev\_make\_7**

**Intercepts for Publisher ID’s for 2 and 3 is calculated as the relative effect as compared to the average effect of all other Publisher IDs. Same is the case for Device\_make IDs.**

1. Plot the ROC curve for this model, and report the area under the ROC curve.





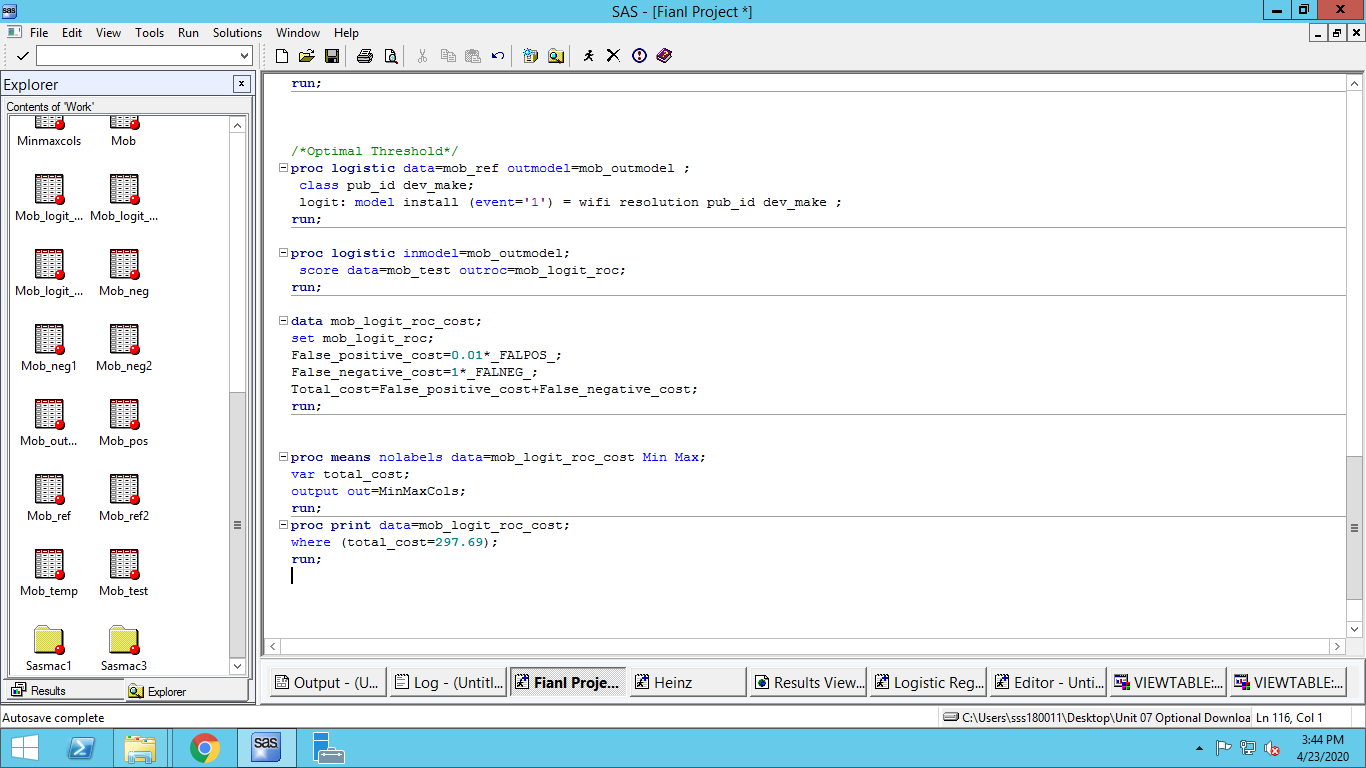
**Area under the curve: 0.5953**

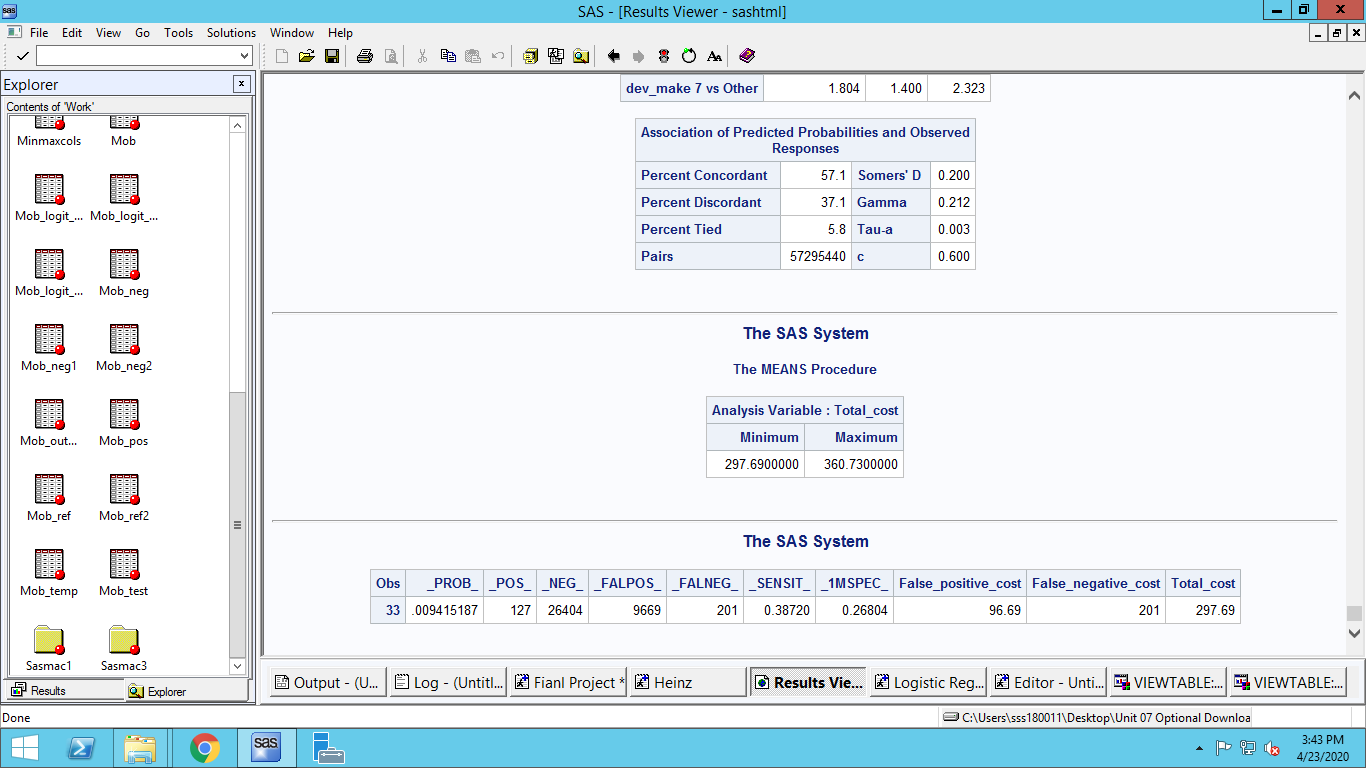
**The logistic regression ROC curve was created on test data**

1. The advertising platform would like to determine whether to show the ad from this advertiser depending on the publisher and consumer characteristics. In particular, the advertising platform needs to come up with a threshold such that if the probability of installing the ad is above that threshold, the ad is shown to the consumer.

Showing an ad to a consumer who would not install the app results in some inconvenience cost to the consumer which in turn leads to less participation and causes a loss of 1 cent to the platform. On the other hand, not showing an ad to a consumer who would have installed the app results in a missed opportunity cost of 100 cents to the platform. The platform would like to minimize the total expected cost. Using the ROC table generated by SAS, plot the total cost for different threshold values. Calculate the threshold at which the cost is minimized and report the cost at this threshold.

1. **First OUTMODEL is created for the logistic model using proc logistic**
2. **Then ROC table is created from the test data using the outmodel of step 1.**
3. **Total cost is calculated for each threshold values by taking number of false positives and negatives from the ROC table and multiplying with cost provided for each type of error.**
4. **Minimum total cost and corresponding threshold is then displayed**



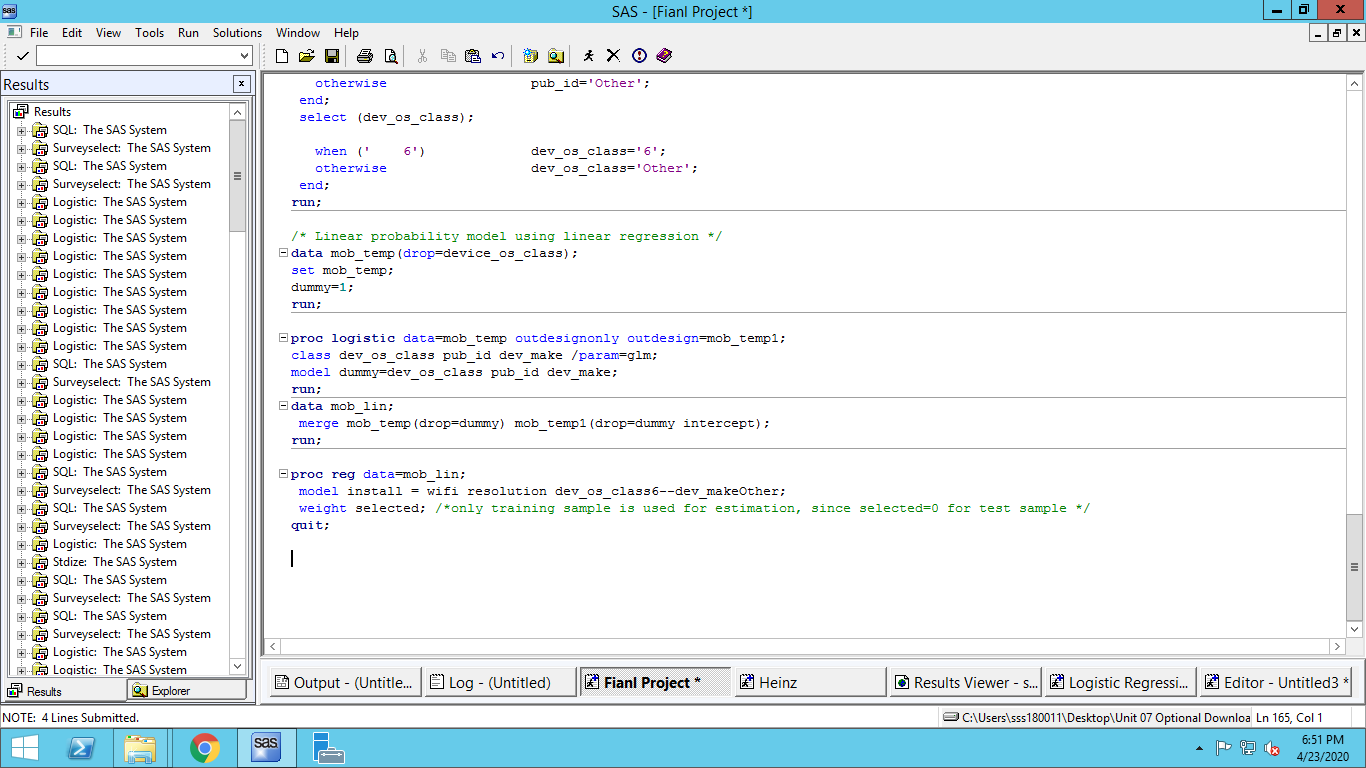


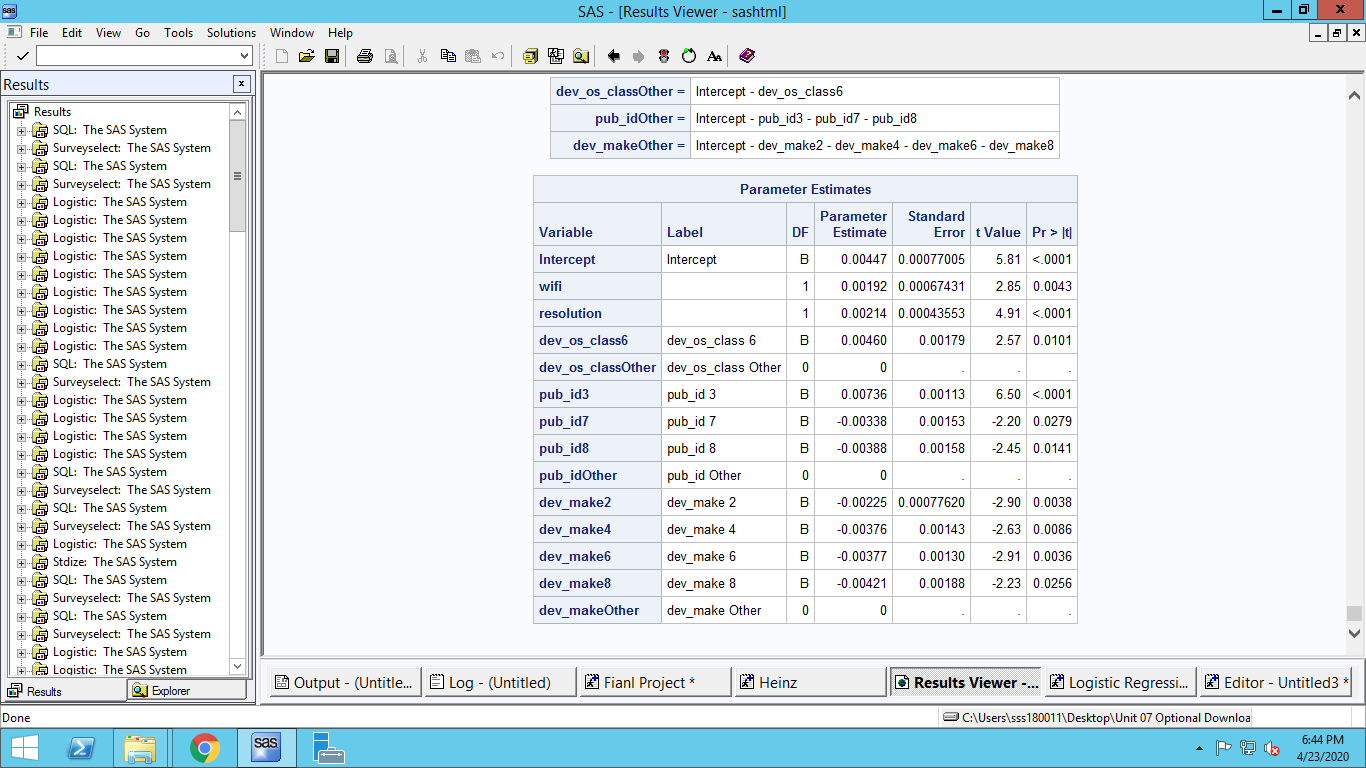
**Threshold: 0.009415187**

**Cost: $297.69**

# Part II. Linear Probability Model

1. Develop a **linear probability model** to estimate the probability of installing the ad based on publisher and consumer characteristics. Present only the final model and explain the procedure and different measures you have used to come up with this model.





**Data is formatted and categorical values are imputed in place of numerical values for device\_make, publisher\_id and device\_OS\_class.**

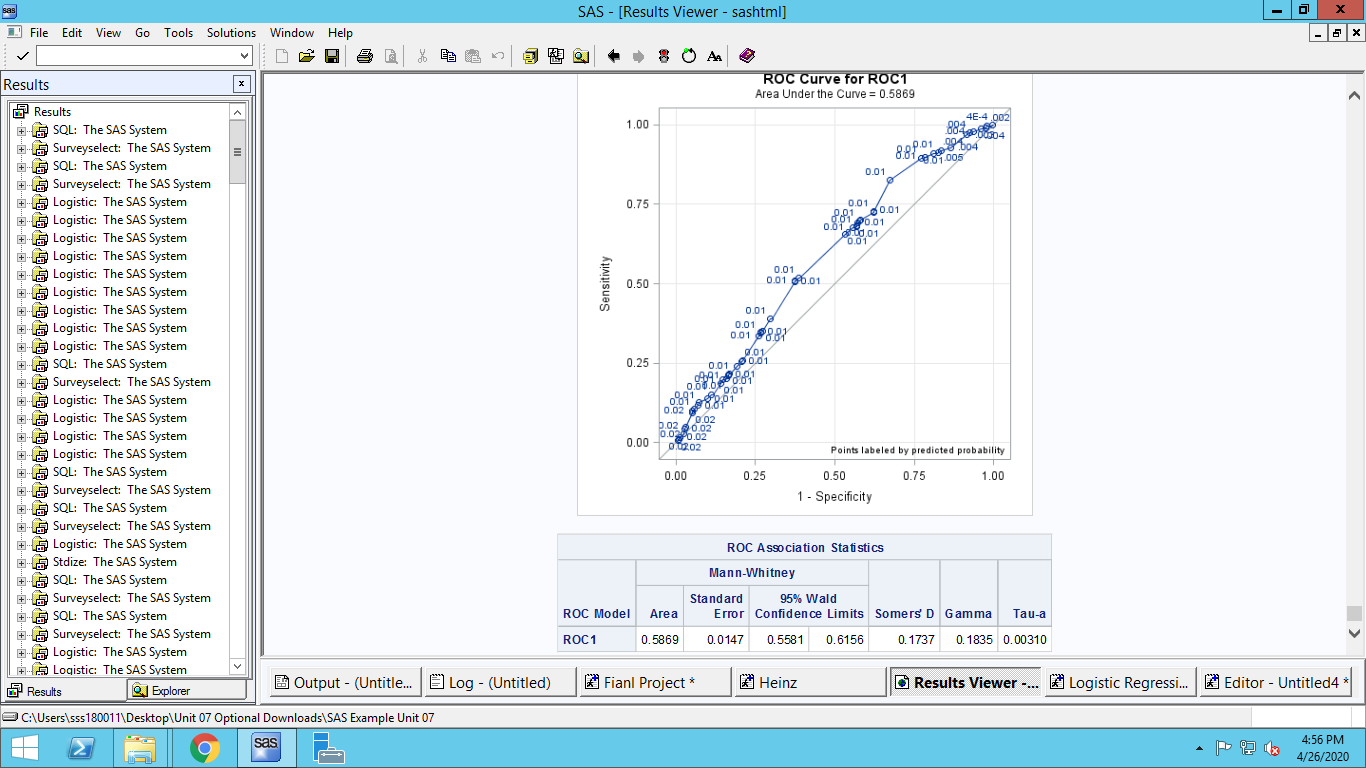
**Volume, device\_width, device\_length, platform\_class were all found to be statistically insignificant.**

**Out of device\_make only values of 2, 4, 6 and 8 is found to be statistically significant.**

**Similarly out of publisher ID only ids’ of 3, 7, 8; and; out of device\_OS\_class only class 6 is statistically significant.**

**THE MODEL: Install = 0.00447 \* Intercept + 0.00192 \* wifi + 0.00214 \* resolution + 0.00460 \* dev\_os\_class6 + 0.00736 \* pub\_id3 + -0.00338 \* pub\_id7 + -0.00388 \* pub\_id8 + -0.00225 \* dev\_make2 + -0.00376 \* dev\_make4 + -0.00377 \* dev\_make6 + -0.00421 \* dev\_make8**

1. Plot the ROC curve for this model and report the area under the ROC curve. Note that in order to this, you must use the *Proc logistic* command, but without fitting the model. (Refer to the lecture). At the 95% confidence level, is the area below this graph higher than the one you obtained from the logistics regression? (Hint: You need to look at the confidence intervals)
2. **Predictions are made for test data using proc reg**
3. **Plot of ROC curve is created using predictions from linear model**



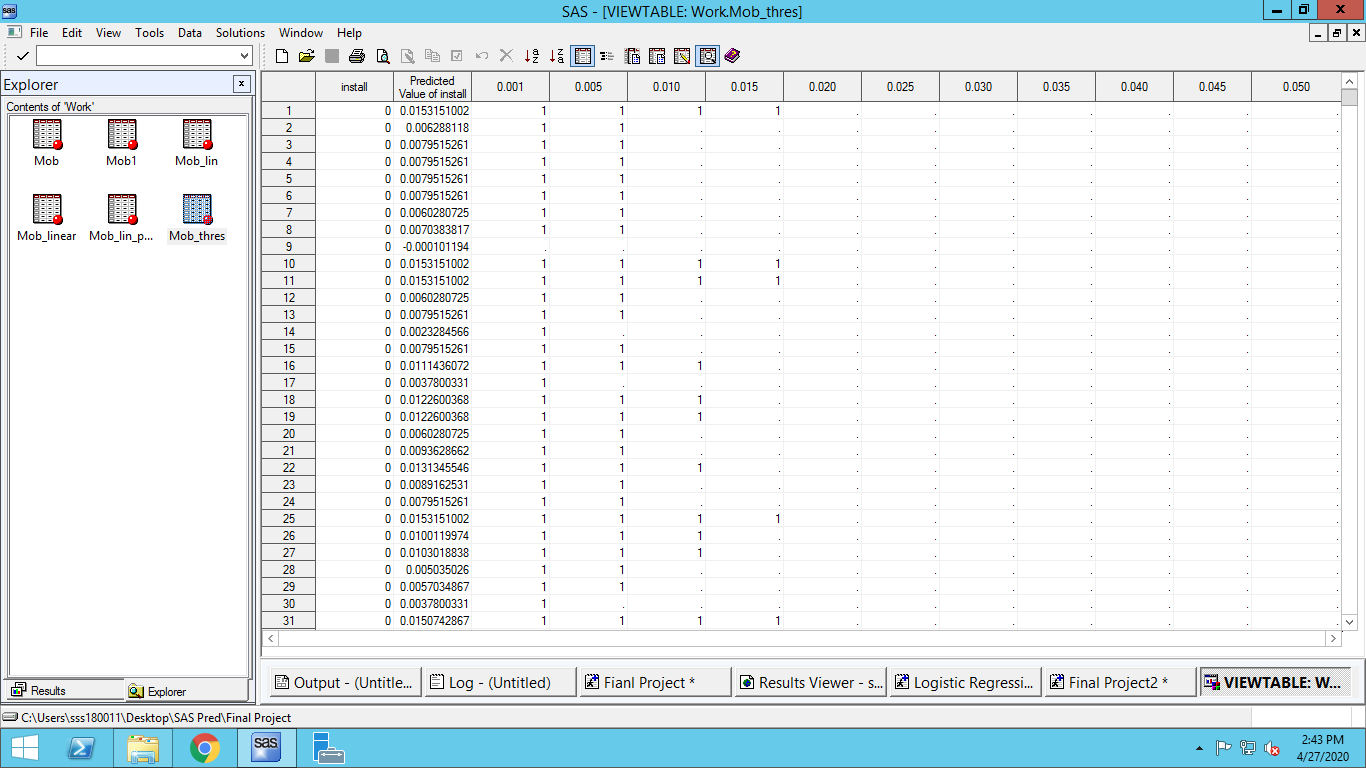
**AREA UNDER THE CURVE: 0.5869**

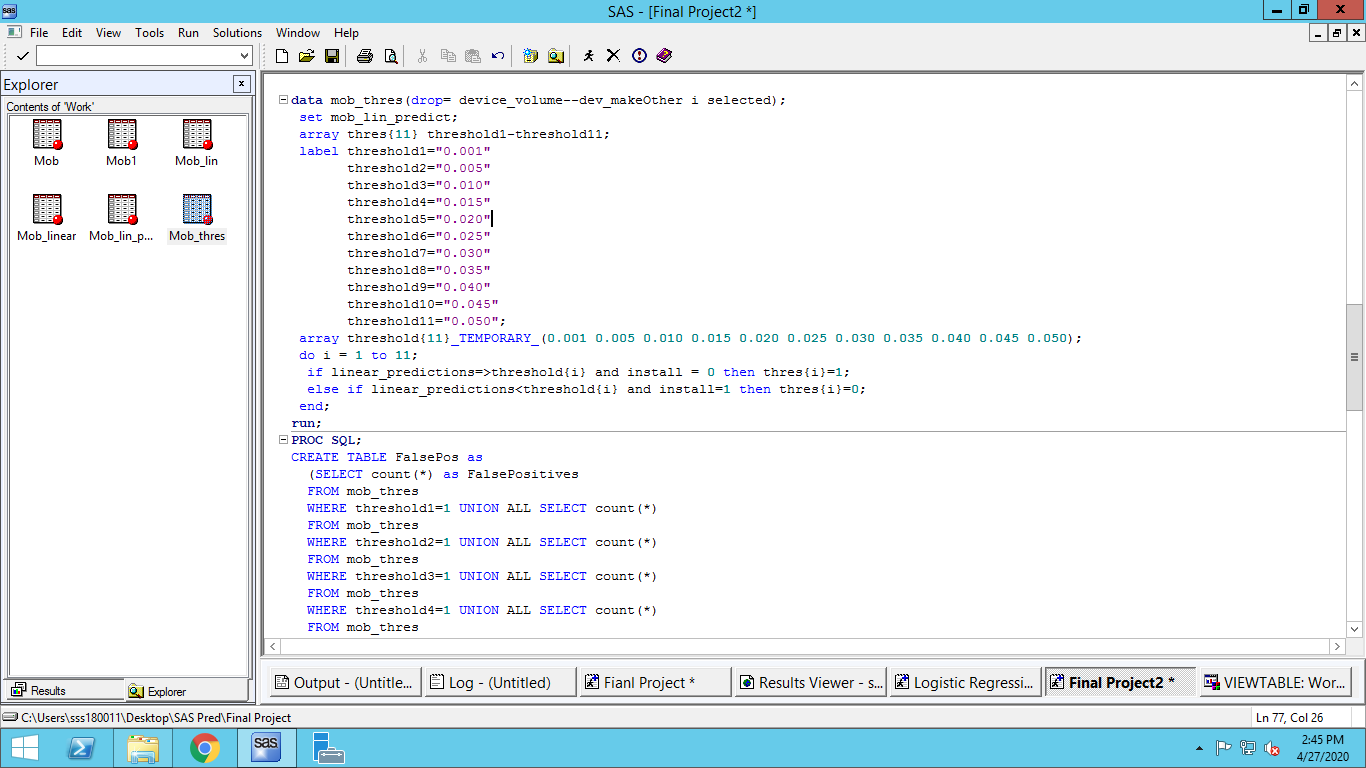
**At 95% Confidence Interval of 0.5581** **and 0.6156** **we can’t say area under the curve is higher than the previous method.**

1. Repeat the analysis you have done in part I to determine the optimal threshold assuming the same numbers for costs. Is the optimal threshold different in two cases? How about the optimal expected cost?

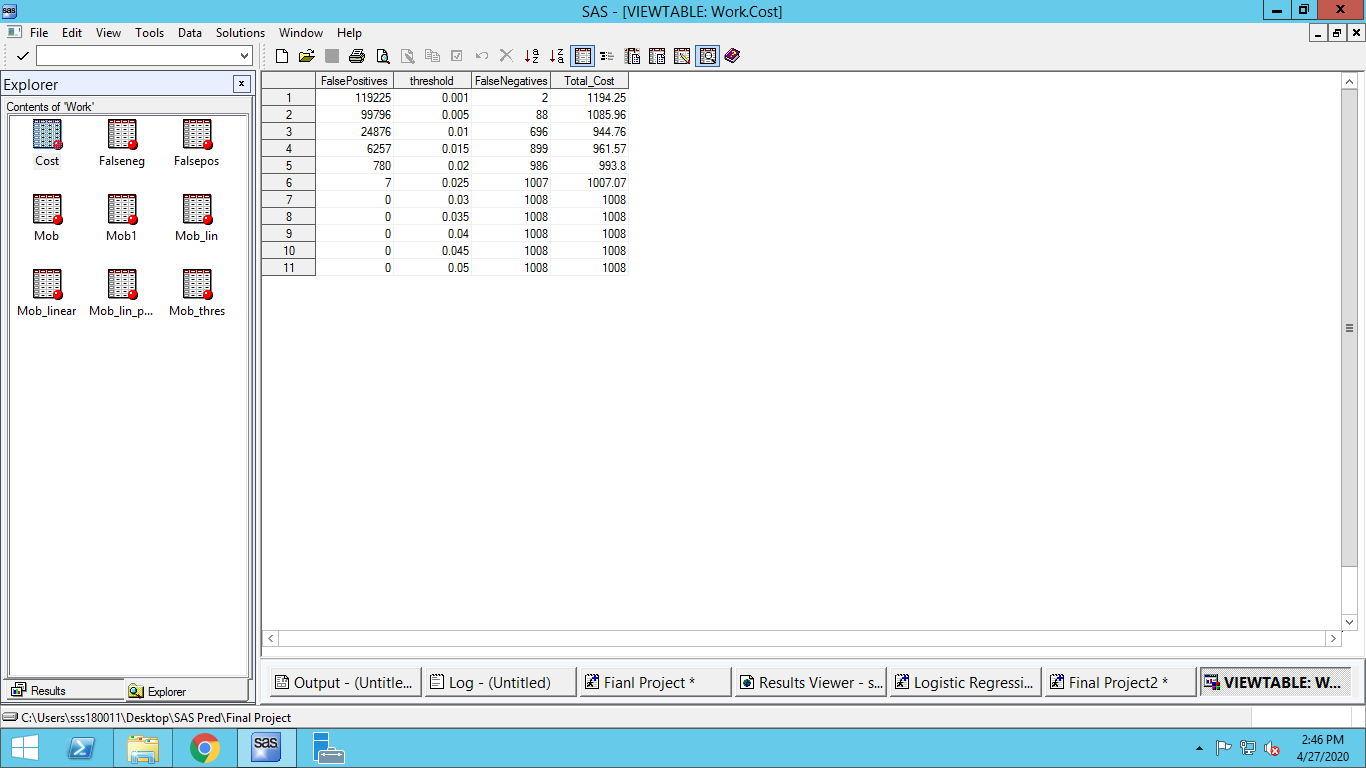
Note that unlike the logistic regression case, SAS does not generate the ROC table automatically. To make your job easier, you can calculate the total cost at these thresholds:

0.001 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045 0.050

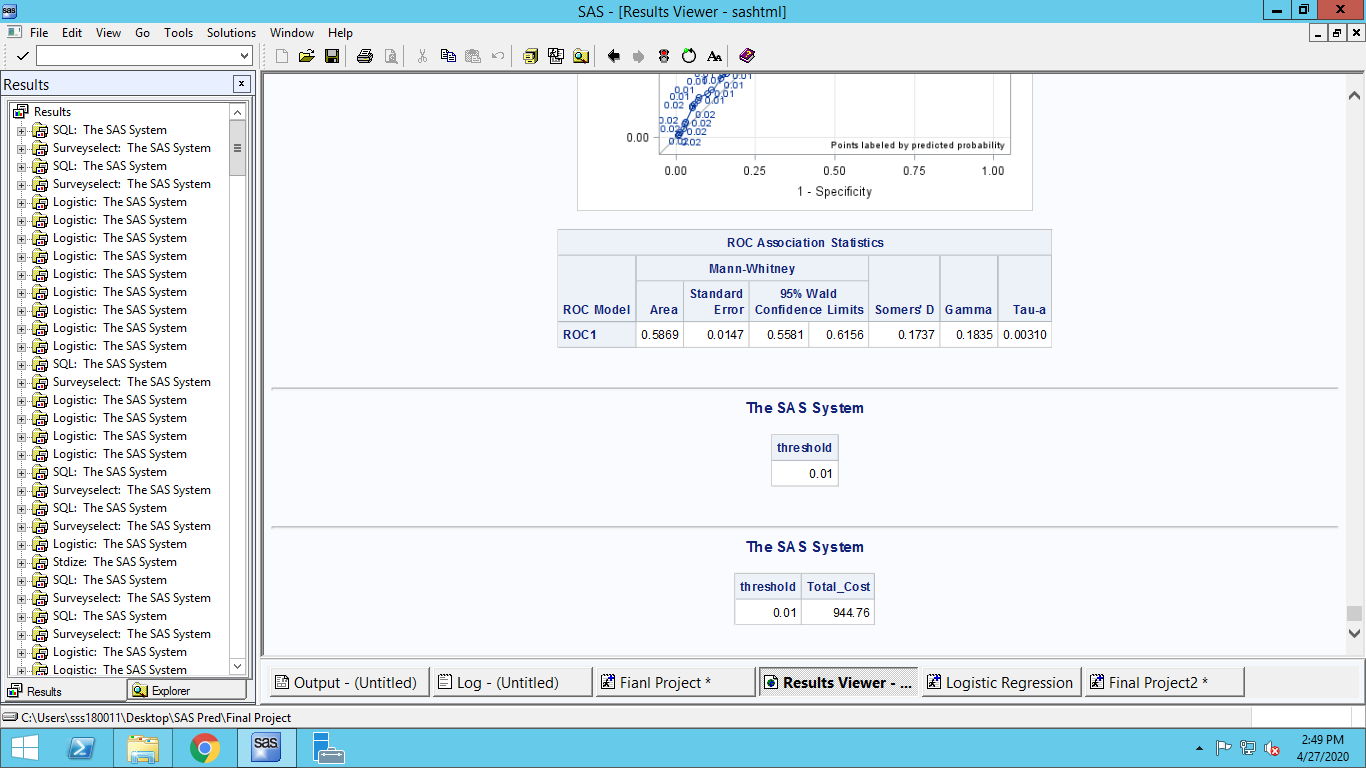




**A dataset was created with all 11 threshold values, having 1 for FalsePositives , 0 for FasleNegatives and missing values for True negatives and positives.**



**The number of false positives and negatives were calculated for each threshold using PROC SQL and combined using JOIN and total cost calculated.**



**The minimum total cost is found to be $944.76 for threshold of 0.01.**

Deliverables

* Project Report: For each question above, describe the model building and selection process that you followed, along with suitable tables and graphs as necessary.
* SAS code: Include a SAS file with detailed comments to reproduce all the results, tables and figures in the report. The code must be clearly labeled so that it is straightforward to see how to reproduce a particular result / table / figure. If the code will not execute, then points will be deducted. The code should assume that it will be executed in the folder containing the dataset.