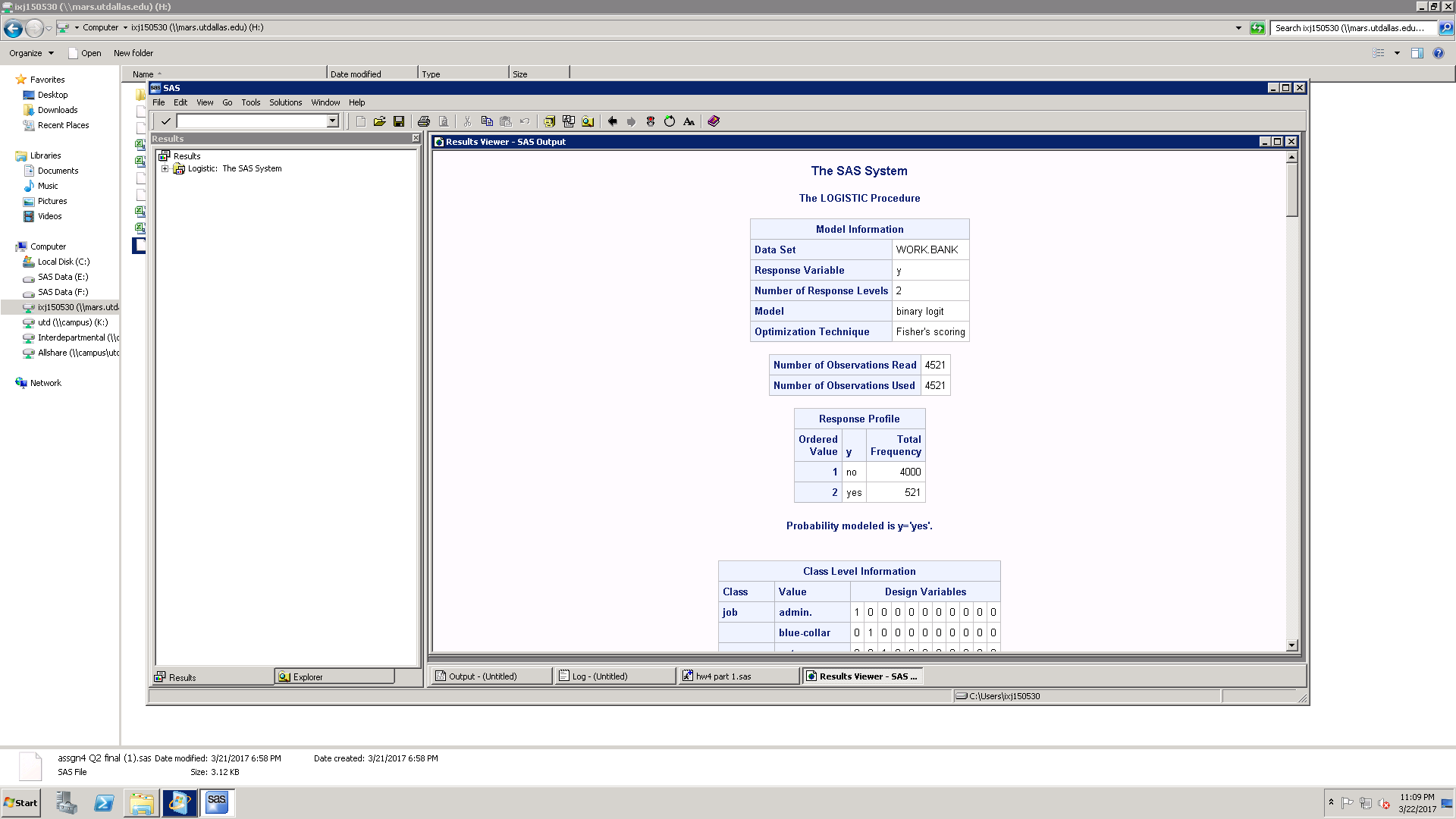
**Assignment: Logistic Regression**

Q1. The following data is from a direct marketing campaign of a bank. The goal of the campaign was to get the customer to subscribe a term deposit. The marketing campaign consisted of making multiple phone calls to the customer. The data is in file: bank.csv (randomly selected, 4521 observations).

I built a predictive modelto predict if the client will subscribe a term deposit (variable y).

Used Logistic regression to build a model.



What percent of customers in data subscribed to a term deposit?

Ans: Total number of observations = 4521

Total Observations where people subscribed to term deposit = 521

Percent of customers subscribed to term deposit= 521/4521=11%

Meaning of each (significant) coefficient:

Ans: The significant variables are:

* Tertiary education
* Loan = yes
* Contact = cellular
* Contact = telephone
* Day
* Month = apr
* Month = Aug
* Month = Feb
* Month = June
* Month = Mar
* Month = Oct
* Month = September
* Duration
* Campaign
* Poutcome = success

Interpretation:

1. Education = Tertiary: We can say that when the education is tertiary the percentage of subscribing to a term deposit increases by 109% as compared to unknown the percent is much less.

2. Loan = yes: If the person has a loan then percentage of subscribing to a term deposit decreases by 47% as compared to when the loan is not taken.

3. Contact = cellular: When the contact communication type is cellular as compared to unknown then percentage of subscribing to a term deposit increases by 312%.

4. Contact = telephone: When the contact communication type is cellular as compared to unknown then percentage of subscribing to a term deposit increases by 284%.

5. Day: The percentage of subscribing to a term deposit increases by 1% if there is a change in the last day of the contact by 1 day.

6. Month = apr: If the month is April and not Jan then the percentage of subscribing to a term deposit increases by 207%.

7. Month = Aug: If the month is August and not Jan then the percentage of subscribing to a term deposit increases by 125%.

8. Month = Feb: If the month is August and not Jan then the percentage of subscribing to a term deposit increases by 276%.

9. Month = June: If the month is June and not Jan then the percentage of subscribing to a term deposit increases by 435%.

10. Month = Mar: If the month is March then the odds of subscribing to a term deposit is 13.763 times more than jan.

11. Month=Oct: If the month is October then the odds of subscribing to a term deposit is 11.957 times more than jan.

12. Month = September: If the month is September and not Jan then the percentage of subscribing to a term deposit increases by 493%.

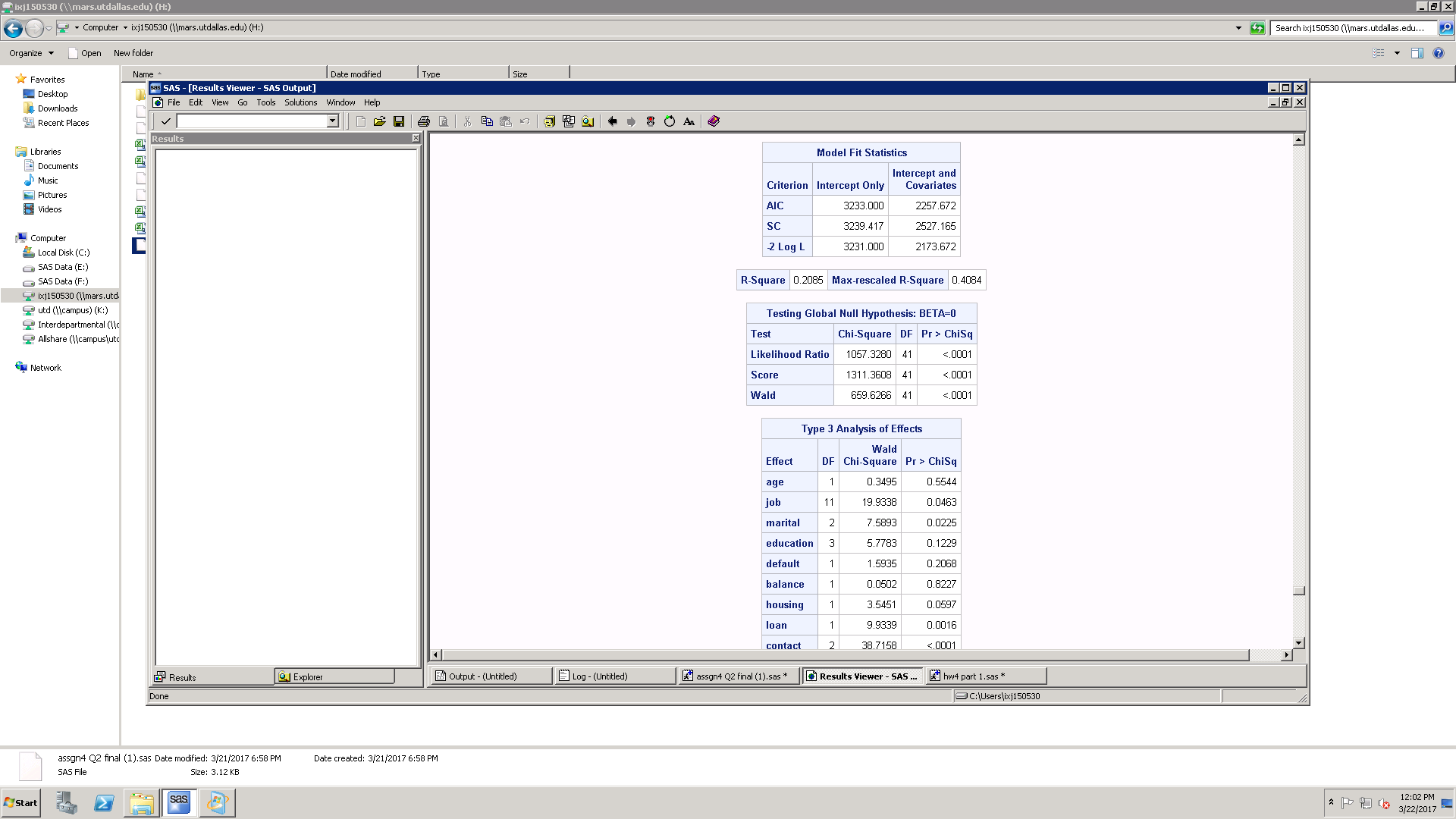
13. Duration: If there is a second change in duration then the odds of subscribing to a term deposit is 1.004 times more.

14. Campaign: The percentage of subscribing to a term deposit decreases by 7% if there is a change in the last day of the contact by 1 day during the campaign.

15. Poutcome = success: if the outcome of the previous marketing campaign is success then the odds of subscribing to a term deposit is 12.777 times more than is the outcome is unknown.

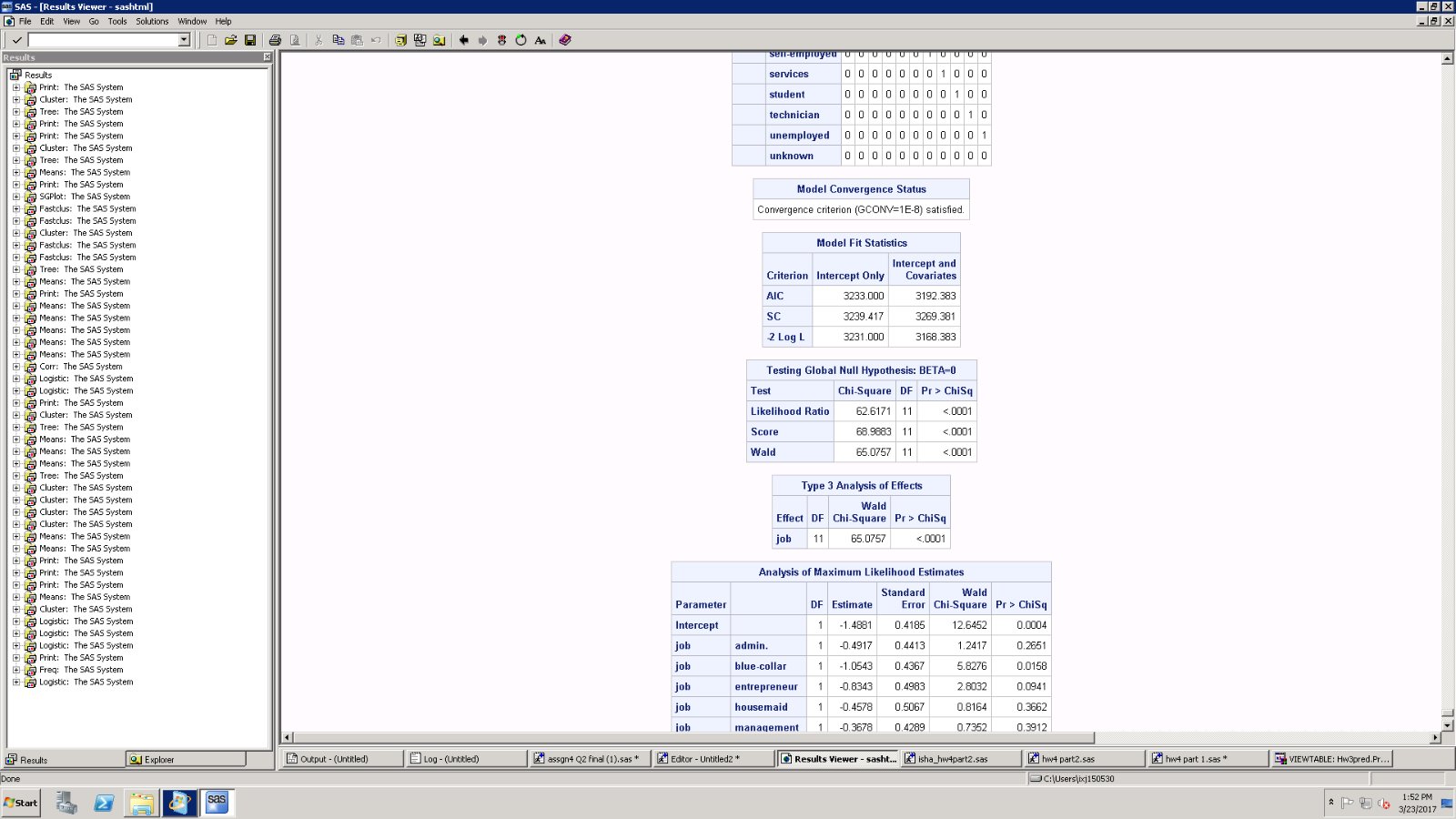
Interpreted the fit of the model based on -2logL, AIC, SC.

For a model to be fit the values -2logL, AIC and SC should be as small as possible.



If I compare the intercept and covariates model with the intercept only model, there is a significant difference between the two values and the values of model with covariates is smaller than the intercept model. Therefore, we can say that this model is a good fit with covariates than without them.

Now if I run logistic with only one covariate I get the following result.

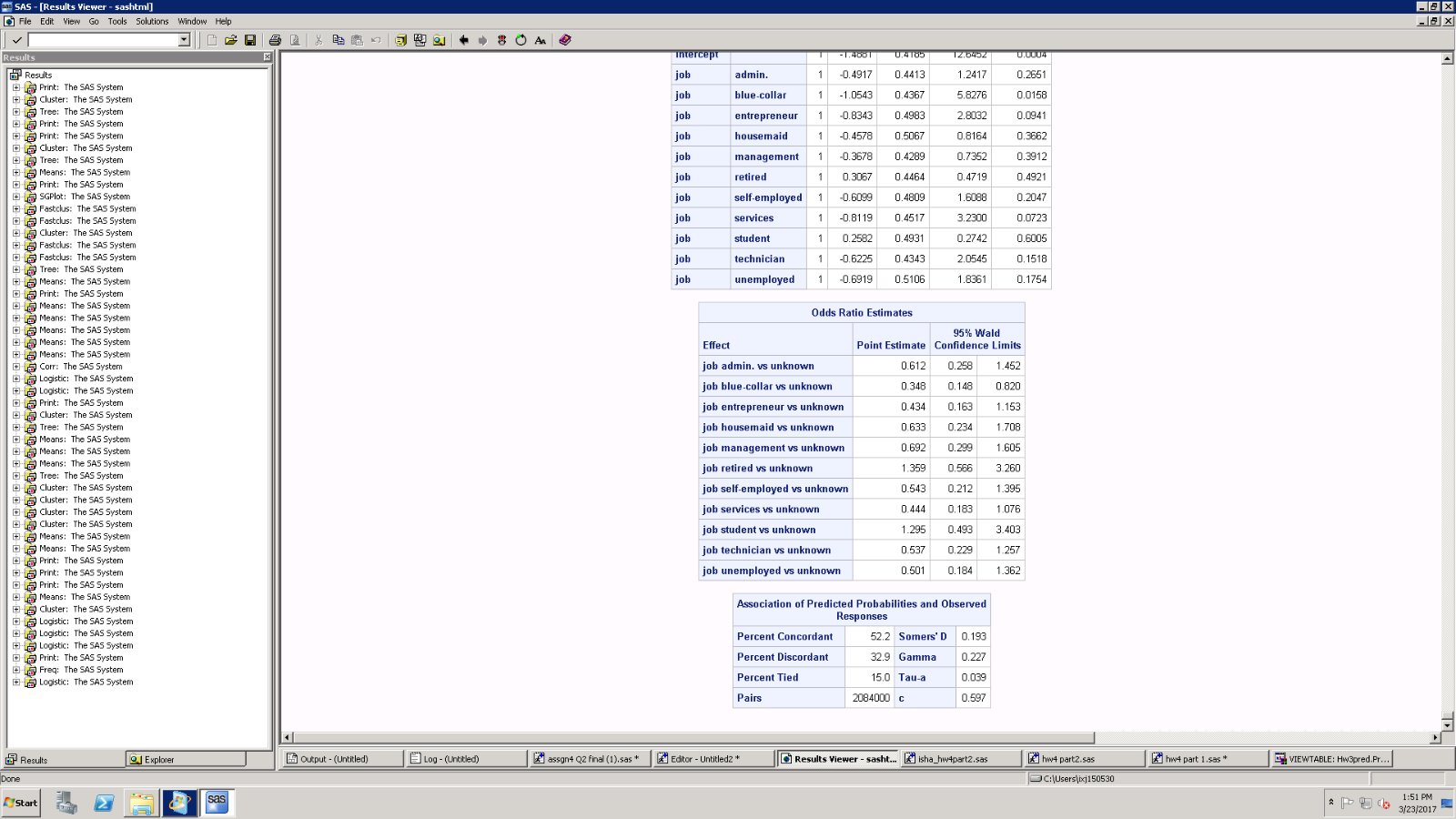


This model is much worse than the model with all the covariates as there is not much change is the AIC, SC and -2LogL values and they are too high.

Concluding after comparison to other models this model is a good fit.

What is the percent concordant? What does it mean?

**Ans: -** A pair is said to be concordant if the observation with the higher response also has the higher estimated probability the numbers given are the percentages of pairs in each of these classes; obviously, the higher the percentage of concordant pairs the better is the fit of the model. Thus, this model is a good fit as indicated by percent concordant (90.1%).

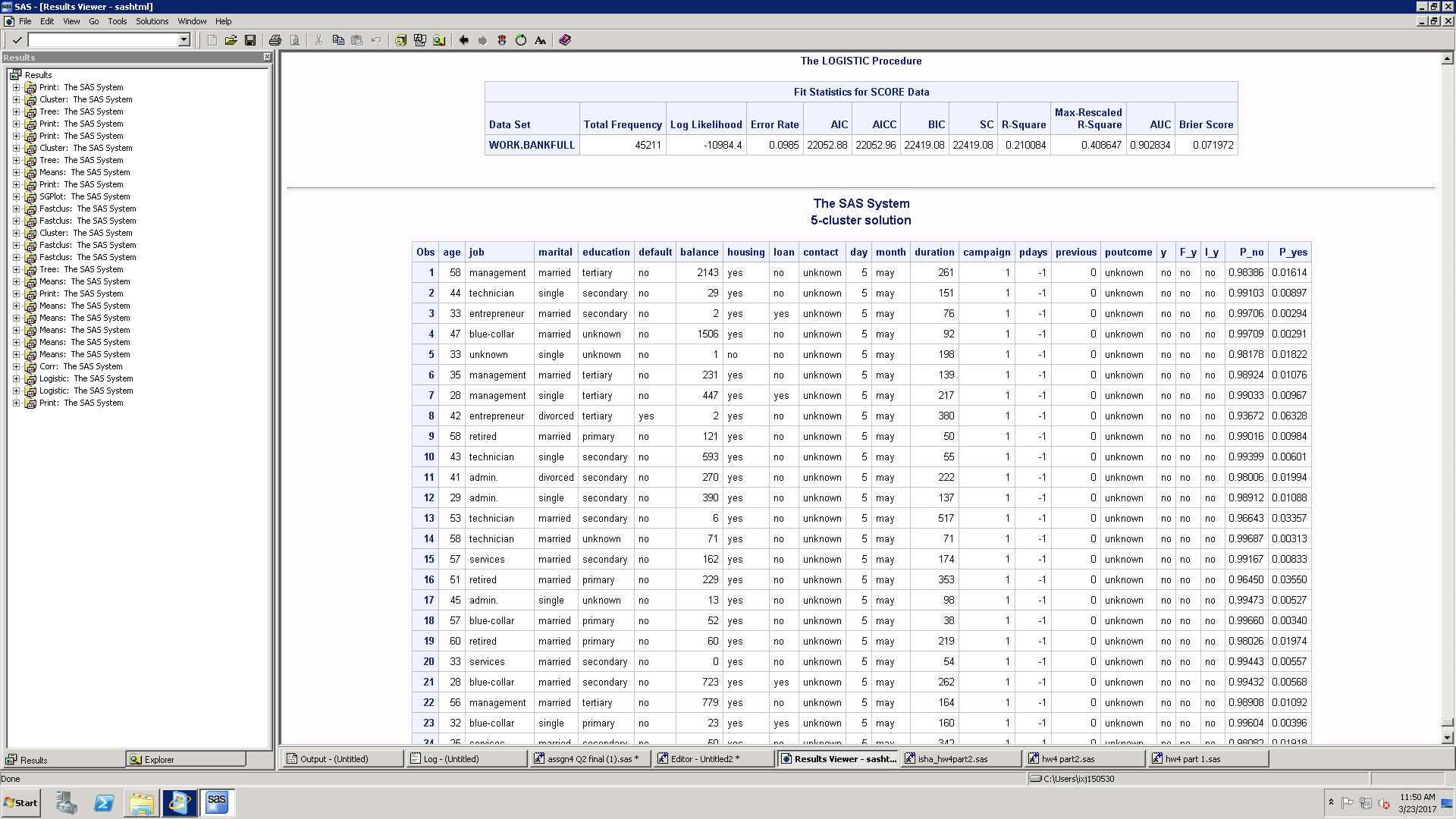


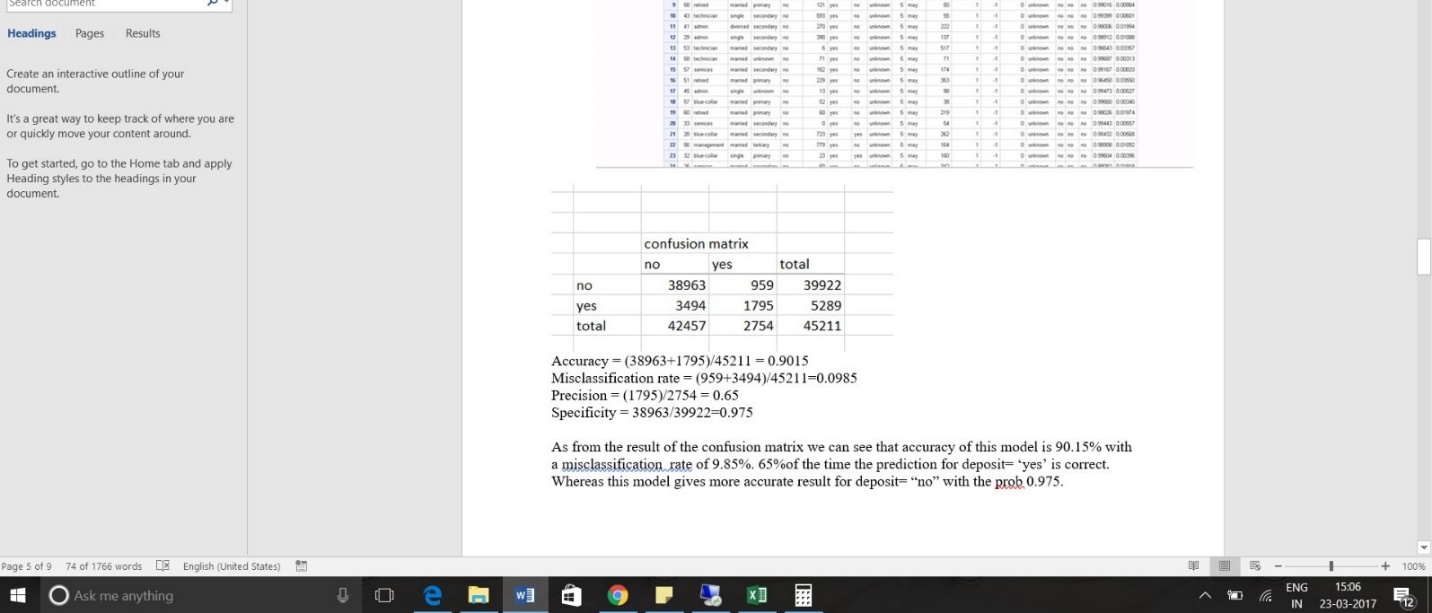
If I run the null model and look at the percent concordant it is very low that is 52%.

Therefore, when we compare we can say that the current model is good.

Used the parameters to predict whether a customer will subscribe to a term deposit for the new data bank-full.csv:

If we see the proc score result below:





Accuracy = (38963+1795)/45211 = 0.9015

Misclassification rate = (959+3494)/45211=0.0985

Precision = (1795)/2754 = 0.65

Specificity = 38963/39922=0.975

As from the result of the confusion matrix we can see that accuracy of this model is 90.15% with a misclassification rate of 9.85%. 65% of the time the prediction for deposit= ‘yes’ is correct. Whereas this model gives more accurate result for deposit= “no” with the prob 0.975.