


AMES, IOWA HOUSING DATASET ANALYSIS

Research Question: What variables are the strongest predictors for whether or not there is central air in a house?

Siddharth Ahuja

STAT 448 Group 5
Group Members: Rongqi Pan, Sooeun Kim,
Ari Ben-Zeev



Introduction

Over the course of the Fall 2018 semester, our group has been analyzing a dataset that extensively describes several houses in the region of Ames, Iowa. The ‘Ames Housing Dataset’ was compiled by Dean De Cock, who has made the dataset publicly available on Kaggle.

The dataset contains 2930 observations and 83 variables. Each observation represents one house in the Ames region. Some of the important variables in the dataset include ‘SalePrice’ (the property’s sale price in dollars), ‘YearBuilt’ (the year the property was built) and ‘OverallCond’ (a rating of the overall condition of a property ranked with values from 1 to 9).

Based on this data, the research question I decided to focus on was the following: What variables are the strongest predictors for whether or not there is central air in a house? I decided to focus on this question because prospective home buyers in the region may find it important to have central air in their future home. With the analysis in the rest of this report these prospective buyers can see if other features that they are looking for in a house usually correlate with the house having central air.

The variable of concern for this question is appropriately titled ‘CentralAir’ in the dataset (this variable will act as the response for the analysis that will be done in the rest of this report). CentralAir is a binary, categorical variable with values stored as ‘Y’ (signifying there is central air in the house) and ‘N’ (signifying there is no central air in the house). The ‘CentralAir’ variable and this dataset will be analyzed using logistic regression and discriminant analysis. To get a better idea about the dataset we will be working with in the following analysis, basic descriptive statistical results have been provided in ‘Figure 1’ below for the ‘CentralAir’ variable.

CentralAir	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
N	95	YearBuilt	95	1927.57	21.05	1872.00	1970.00
		OverallCond	95	5.07	1.46	1.00	8.00
		SalePrice	95	105264.07	40671.27	34900.00	265979.00
Y	1365	YearBuilt	1365	1974.31	28.34	1880.00	2010.00
		OverallCond	1365	5.61	1.08	2.00	9.00
		SalePrice	1365	186186.71	78805.21	52000.00	755000.00

Figure 1: Basic Descriptive Statistics

Methods and Results for Logistic Regression

There was no data cleaning required once the dataset was inputted into SAS. As mentioned in the introduction, the analysis to answer the research question will involve logistic regression and discriminant analysis. First, logistic regression will be explored.

In order to initially explore the ‘CentralAir’ variable for the purposes of logistic regression, a logistic regression model with all other variables in the dataset (except ‘Id’, which was just a random integer number associated with each house) as predictors was created. Looking at the Cbar plot for this model it was clear that there were several problems with this model. The Cbar values were large with one influential point having a Cbar value of over 8,000,000. This Cbar plot is saved in the appendix as ‘Figure 2’. Attempt at removing the influential points in this model were not successful.

Next, a logistic regression model with just the continuous variables in the dataset as predictors was created. This model had 36 variables as predictors for ‘CentralAir’. With this new model a much more manageable Cbar plot was created. This Cbar plot can be seen in ‘Figure 3’ below.

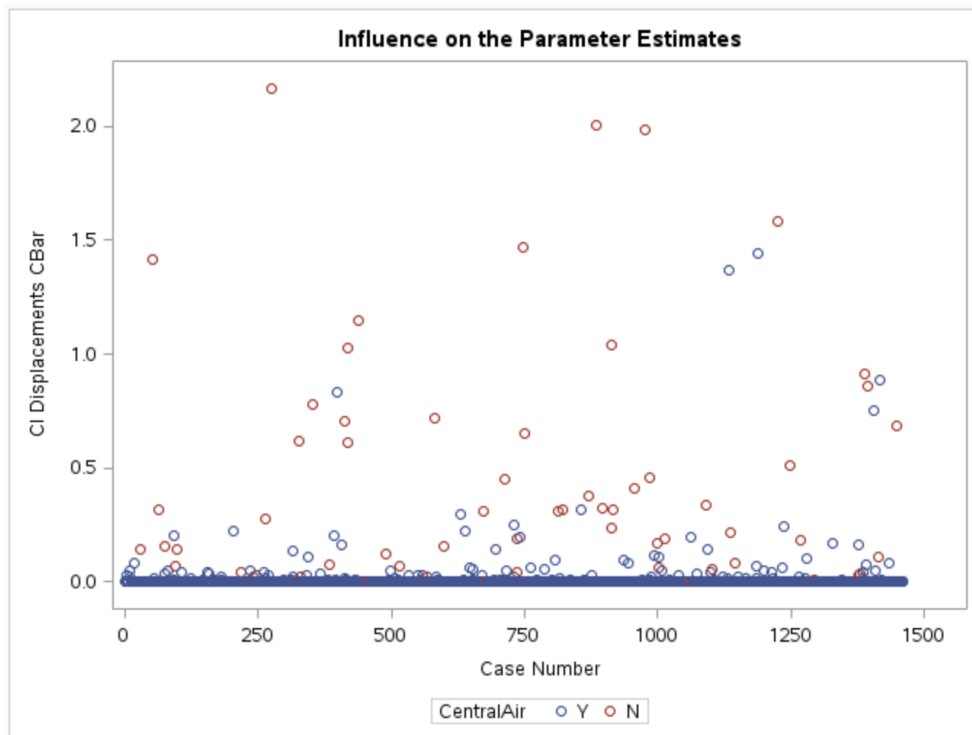


Figure 3: Cbar plot for logistic regression model with only continuous predictors

From Figure 3 we can see that there are several points above a Cbar value of 1, and there are no points that are significantly far away from the rest of the points. Since the points are so well spread out, this suggests that attempting to remove the points with the largest Cbar values will probably not result in a significantly better model. This was indeed the case, as removing any points did not improve the Cbar plot.

After looking at the significance of the parameter estimates for this model, we can see that there are 5 significant predictors. These predictors are ‘YearBuilt’, ‘BsmtUnfSF’ (surface area in square feet of unfinished basement), ‘Fireplaces’ (number of fireplaces), ‘GarageYrBlt’ (year the garage was built) and ‘OpenPorchSF’ (surface area in square feet of open porch). The significance of parameter estimates have been displayed in ‘Figure 4’ on the right.

Next, stepwise selection was performed to determine which variables should be kept in the final model. The 9 variables that were selected by this stepwise selection are displayed in ‘Figure 5’ below.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-330.2	279.4	1.3965	0.2373
MSSubClass	1	-0.00877	0.00512	2.9264	0.0871
LotFrontage	1	0.0234	0.0138	2.8666	0.0904
LotArea	1	-0.00004	0.000028	1.6841	0.1944
OverallQual	1	-0.0597	0.2418	0.0609	0.8051
OverallCond	1	0.6461	0.2068	9.7658	0.0018
YearBuilt	1	0.0560	0.0129	18.9454	<.0001
YearRemodAdd	1	0.0176	0.0114	2.3613	0.1244
BsmtFinSF1	1	0.000953	0.000907	1.1034	0.2935
BsmtFinSF2	1	-0.00077	0.00140	0.3050	0.5808
BsmtUnfSF	1	0.00151	0.000712	4.5065	0.0338
TotalBsmtSF	0	0	.	.	.
FirstFlrSF	1	-0.00086	0.00130	0.4357	0.5092
SecondFlrSF	1	-0.00104	0.00118	0.7776	0.3779
LowQualFinSF	1	-0.00022	0.00295	0.0056	0.9406
GrLivArea	0	0	.	.	.
BsmtFullBath	1	0.1958	0.5269	0.1382	0.7101
BsmtHalfBath	1	0.1894	1.0227	0.0343	0.8531
FullBath	1	0.5450	0.6152	0.7850	0.3756
HalfBath	1	0.2938	0.5719	0.2638	0.6075
BedroomAbvGr	1	-0.3663	0.3572	1.0516	0.3051
KitchenAbvGr	1	-1.5134	0.8221	3.3885	0.0657
TotRmsAbvGr	1	0.4892	0.2597	3.5471	0.0597
Fireplaces	1	1.1275	0.4665	5.8405	0.0157
GarageYrBlt	1	0.0425	0.0112	14.2543	0.0002
GarageCars	1	-0.3959	0.5554	0.5082	0.4759
GarageArea	1	-0.00291	0.00205	2.0103	0.1562
WoodDeckSF	1	7.973E-6	0.00214	0.0000	0.9970
OpenPorchSF	1	-0.00860	0.00299	8.2602	0.0041
EnclosedPorch	1	0.000986	0.00263	0.1402	0.7081
ThreeSsnPorch	1	0.0568	1.1440	0.0025	0.9604
ScreenPorch	1	0.00557	0.00516	1.1693	0.2795
PoolArea	1	0.0241	2.0681	0.0001	0.9907
MiscVal	1	0.000661	0.00155	0.1827	0.6691
MoSold	1	-0.0229	0.0740	0.0959	0.7568
YrSold	1	0.0507	0.1393	0.1325	0.7158
SalePrice	1	0.000010	0.000011	0.9038	0.3418

Figure 4: Parameter Estimates for Logistic Regression Model

Summary of Stepwise Selection							
Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed					
1	YearBuilt		1	1	145.0543		<.0001
2	OverallCond		1	2	46.1993		<.0001
3	Fireplaces		1	3	13.0223		0.0003
4	GarageYrBlt		1	4	12.4112		0.0004
5	MSSubClass		1	5	10.8397		0.0010
6	OpenPorchSF		1	6	5.6267		0.0177
7	TotalBsmtSF		1	7	6.8135		0.0090
8	GarageArea		1	8	4.7500		0.0293
9	SalePrice		1	9	6.2420		0.0125

Figure 5: Variables selected as predictors in stepwise selection

Using the 9 variables selected as predictors in the stepwise selection, I created another logistic regression model. We can see from the Cbar plot of this model in ‘Figure 6’ that there are no highly influential points. From ‘Figure 7’, we can see that the AIC of the new model is lower than the model containing all of the continuous variables.

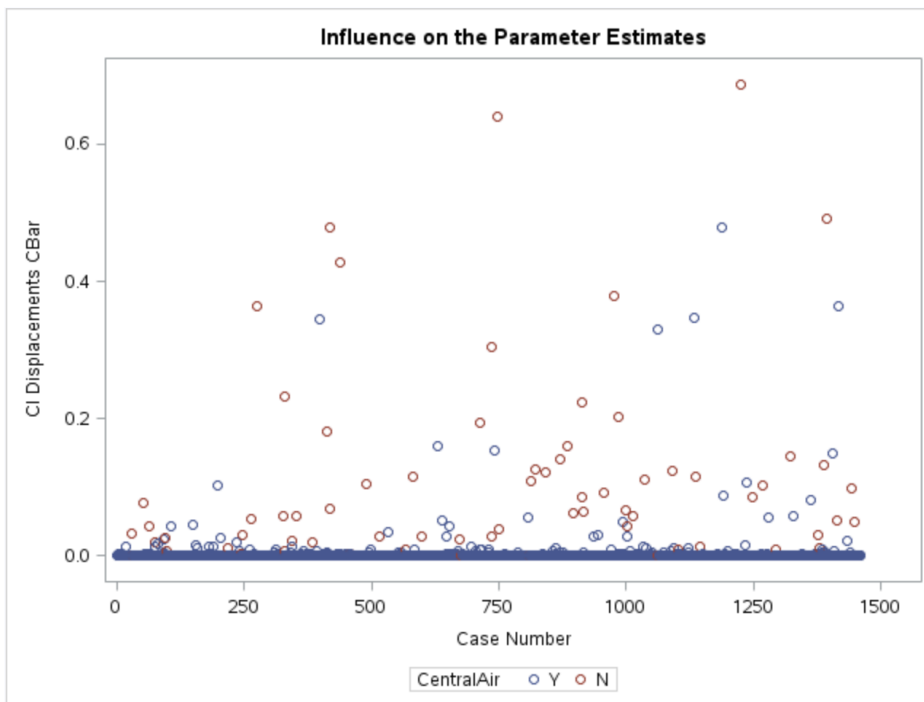


Figure 6: Cbar plot for logistic regression model with only stepwise selected predictors

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	549.799	284.458
SC	555.029	336.749
-2 Log L	547.799	264.458

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
YearBuilt	1.050	1.034	1.068
OverallCond	1.859	1.399	2.470
Fireplaces	2.758	1.274	5.970
GarageYrBlt	1.042	1.024	1.061
MSSubClass	0.989	0.982	0.995
OpenPorchSF	0.993	0.988	0.997
TotalBsmtSF	1.002	1.000	1.003
GarageArea	0.997	0.995	0.999
SalePrice	1.000	1.000	1.000

Figure 7: AIC values and odds ratio estimates for logistic regression model with only stepwise selected predictors

This model was chosen as the final logistic regression model. This was because the 9 variables selected through stepwise selection made more intuitive sense than the 5 variables selected by looking at the significance of the parameter estimates. The significance of parameter estimates model included variables which contained information like the surface area in square feet of unfinished basement. Logically, such variables should not be correlated with whether or not the house has central air.

Lastly, I looked at the odds ratio estimates of the variables in the final model. These are shown in 'Figure 7'. From these odds ratios we can conclude the following (the odds ratios not included in the list below can also be determined by looking at the 'Point Estimate Column' in the 'Odds Ratio Estimates' table):

- For every 1 unit increase in YearBuilt, the model expects the odds of the home having central air to be increased by 5.0%.
- For every 1 unit increase in OverallCond, the model expects the odds of the home having central air to be increased by 85.9%.
- For every 1 unit increase in OpenPorchSF, the model expects the odds of the home having central air to be decreased by 0.7%.

Methods and Results for Discriminant Analysis

For the discriminant analysis part of the analysis, the variables selected by stepwise selection are shown in 'Figure 8' on the next page. Out of the ten variables that were selected, two variables had a partial R-square value greater than 0.05, 'YearBuilt' and 'OverallCond'. This means that each of these two variables account for at least 5% of the variation in the model.

Stepwise Selection Summary										
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	YearBuilt		0.1296	167.46	<.0001	0.87043081	<.0001	0.12956919	<.0001
2	2	OverallCond		0.0894	110.41	<.0001	0.79257401	<.0001	0.20742599	<.0001
3	3	KitchenAbvGr		0.0249	28.72	<.0001	0.77280706	<.0001	0.22719294	<.0001
4	4	Fireplaces		0.0093	10.59	0.0012	0.76558296	<.0001	0.23441704	<.0001
5	5	OpenPorchSF		0.0106	12.03	0.0005	0.75745169	<.0001	0.24254831	<.0001
6	6	MSSubClass		0.0046	5.19	0.0229	0.75395858	<.0001	0.24604142	<.0001
7	7	GarageYrBlt		0.0050	5.67	0.0174	0.75015919	<.0001	0.24984081	<.0001
8	8	GarageCars		0.0094	10.57	0.0012	0.74313637	<.0001	0.25686363	<.0001
9	9	BedroomAbvGr		0.0041	4.58	0.0326	0.74010131	<.0001	0.25989869	<.0001
10	10	FullBath		0.0045	5.03	0.0252	0.73678309	<.0001	0.26321691	<.0001

Figure 8: Variables selected through stepwise selection (variables that explain at least 5% of the variation in the model are highlighted in yellow)

I decided to continue forward with three discriminant analysis models. The first discriminant analysis will contain all the continuous variables in the dataset without stepwise selection (this model will be used as a baseline to compare to the other models), the second analysis will contain all the variables selected by the stepwise selection and the third analysis will only contain the variables which have a partial R-square value of greater than or equal to 0.05.

In the case of all three of these models, the null hypothesis for the test of homogeneity within covariance matrices was rejected, as is shown in 'Figure 9'. Thus, for all three models a quadratic discriminant analysis should be used.

Test of Homogeneity of Within Covariance Matrices	Test of Homogeneity of Within Covariance Matrices	Test of Homogeneity of Within Covariance Matrices																		
<table> <tr> <th>Chi-Square</th><th>DF</th><th>Pr > ChiSq</th></tr> <tr> <td>3115.504263</td><td>666</td><td><.0001</td></tr> </table>	Chi-Square	DF	Pr > ChiSq	3115.504263	666	<.0001	<table> <tr> <th>Chi-Square</th><th>DF</th><th>Pr > ChiSq</th></tr> <tr> <td>548.239259</td><td>55</td><td><.0001</td></tr> </table>	Chi-Square	DF	Pr > ChiSq	548.239259	55	<.0001	<table> <tr> <th>Chi-Square</th><th>DF</th><th>Pr > ChiSq</th></tr> <tr> <td>48.207264</td><td>3</td><td><.0001</td></tr> </table>	Chi-Square	DF	Pr > ChiSq	48.207264	3	<.0001
Chi-Square	DF	Pr > ChiSq																		
3115.504263	666	<.0001																		
Chi-Square	DF	Pr > ChiSq																		
548.239259	55	<.0001																		
Chi-Square	DF	Pr > ChiSq																		
48.207264	3	<.0001																		
Without Selection	With Selection	With Selection and Partial R-Square Greater than 0.05																		

Figure 9: Results of the test of homogeneity within covariance matrices for all three models

The final error rates from the quadratic discriminant analysis of all three models are shown in ‘Figure 10’. The overall error rate is reduced significantly in the models with selection as compared to the model without selection. The model without selection has an overall error rate of 22.98%, the model with selection has an overall error rate of 7.61% and the model with selection and partial R-squares greater than 0.05 has an overall error rate of 6.44%.

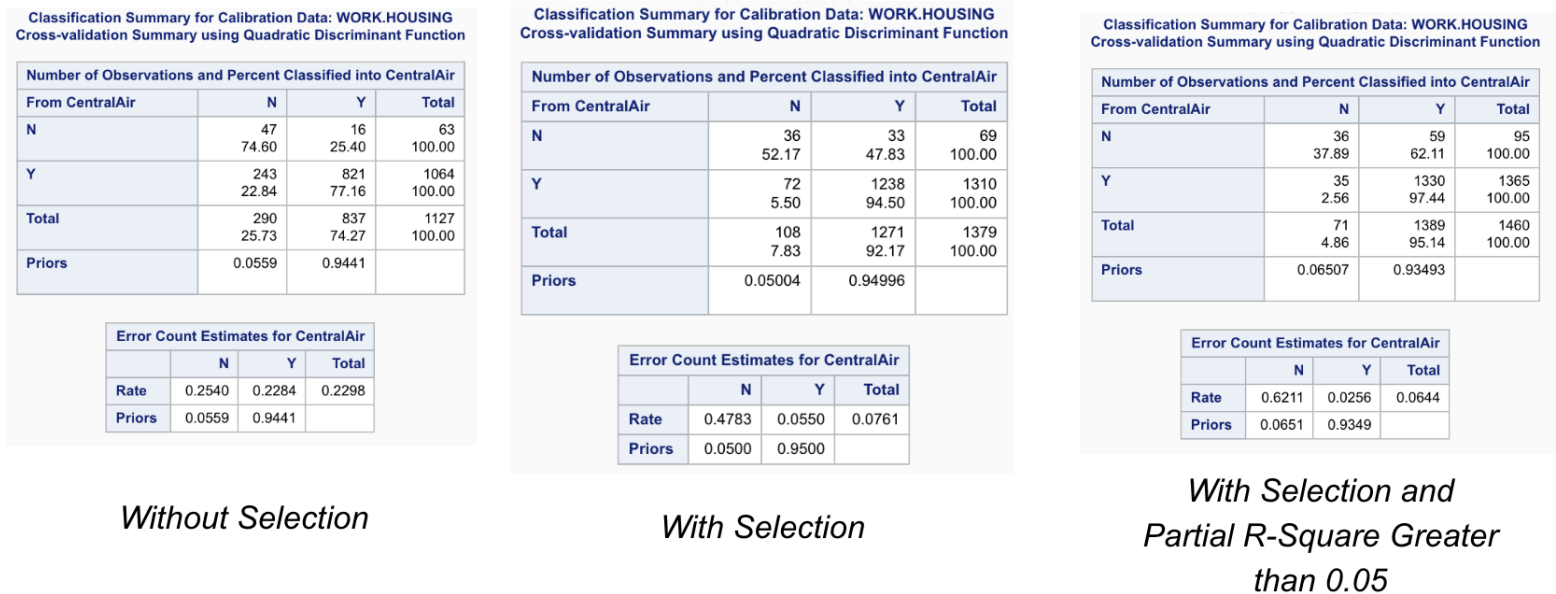


Figure 10: Cross validation summary and error rates for all three discriminant analysis models

At first, it may seem like the models with selection are better in every way compared to the model without selection. However, the error rates for the houses which have no central air have actually increased. This means that a limitation of the models with selection is that more houses without central air are categorized as houses with central air.

Conclusion

The initial purpose of this analysis was to answer the following question: What variables are the strongest predictors for whether or not there is central air in a house? Through this project reasonable final models and results have been found to answer this question.

In the case of the logistic regression model the following variables were found to be good predictors: YearBuilt, OverallCond, Fireplaces, GarageYrBlt, MSSubClass, OpenPorchSF, TotalBsmtSF, GarageArea and SalePrice. The model after the stepwise selection had a lower AIC than the model before the selection (AIC was reduced from 549.799 to 284.458). Thus, we can conclude that the analysis was successful in creating a better model.

In the case of the discriminant analysis, the model with selection and partial R-square values greater than 0.05 did reduce the error rate from 22.98% to 6.44% compared to the model without selection. However, though the overall error rate was improved the percentage of houses without central air misclassified as houses with central air increased from 25.40% to 62.11%. Therefore, we can conclude that this analysis was only partially successful in better classifying the houses with and without central air.

Possible further work would include solving the parts of the project which were not conclusively answered. This included the limitation in the discriminant analysis where the error rate for houses without central air increased as I went further along in the analysis. This could have been improved by using a dataset where there is a better balance of houses with and without central air. From the descriptive statistics in the introduction we can see that the dataset only had 95 houses without central air and 1365 houses with central. For this reason, when houses without central air are misclassified this results in a relatively smaller impact to the total error rate of the discriminant analysis.

Appendix

The following were the additional tests / results found during the course of this project.

This Cbar plot has been discussed in the ‘Methods and Results for Logistic Regression Section’ of the report.

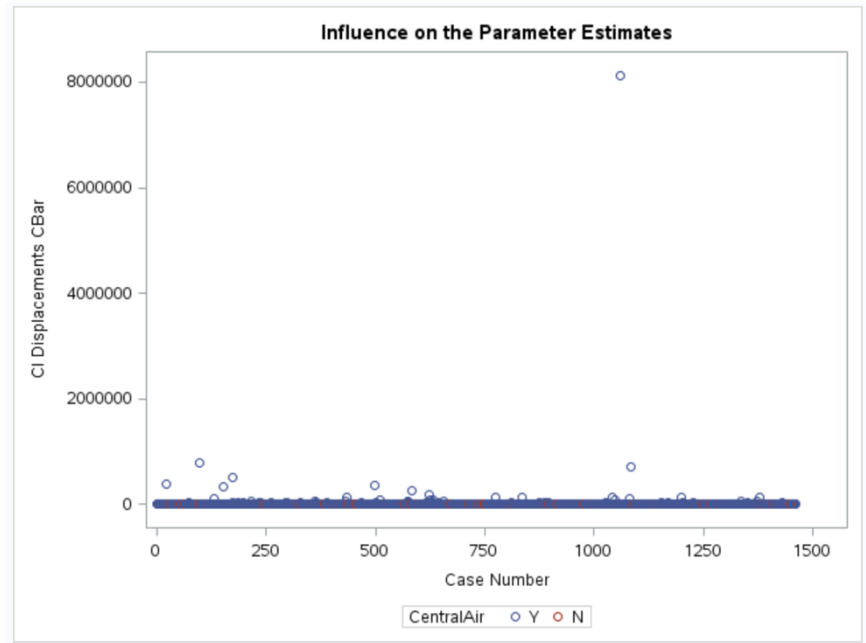


Figure 2: Cbar plot for logistic regression model with both continuous and categorical variables.

For the logistic regression model with only the continuous variables included the global tests were performed. The results from this can be seen in ‘Figure 11’. All the tests reject the null hypothesis, thus at least one of the betas in the model are significantly different from zero.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	261.4891	34	<.0001
Score	305.1445	34	<.0001
Wald	82.7275	34	<.0001

Figure 11: Global Tests for Logistic Regression Model

For the discriminant analysis, the multivariate statistics and exact F statistics table was produced. The null hypothesis for all the tests in this table were rejected. Thus, we can conclude that there is a significant difference across groups.

Multivariate Statistics and Exact F Statistics					
S=1 M=0 N=727.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.77393533	212.79	2	1457	<.0001
Pillai's Trace	0.22606467	212.79	2	1457	<.0001
Hotelling-Lawley Trace	0.29209763	212.79	2	1457	<.0001
Roy's Greatest Root	0.29209763	212.79	2	1457	<.0001

*MultStat for case with Selection and Partial
R-Square Greater than 0.05*

References

The dataset used for this analysis can be found at this link:

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>