Prasanna Krishna Rao

I think I must get 60 Bingo points as I believe I have performed all these:

- (20) Developed a LOGISTIC / POISSON model (if you like it, you may select this as your champion)
- (20 Points) Used decision tree software impute missing
- (10 Points) Used SAS Macros or use, in my opinion, good programming technique
- (10 Points) Handed in your SCORED FILE as a SAS DATA SET and save me to trouble of converting it.

WINE SALES PROJECT

The main intention of this assignment is to predict the number of wine bottles that could be sold or bought using a set of 14 variables. Since count data is to be predicted the following models would be built

1) Poisson Regression
2) Negative Binomial Regression
3) Zero Inflated Poisson Regression
4) Zero Inflated Negative Binomial Regression
5) Linear REGRESSION
6) Logistic Hurdle Regression

Each model would be assessed depending upon their AIC values and each variable chosen from the model would depend upon the p value. The final model would be the ensemble model which would be the average of these models.

1. DATA EXPLORATION (40 points)

The number of records were 12795 with 13 variables apart from the target variable.

A) Mean , Standard Deviation and Median

Using Proc means, Mean, Standard Deviation and median were calculated.

Variable	N Miss	Mean	Median	Std Dev
INDEX	0	8069.98	8110.00	4656.91
TARGET	0	3.0290739	3.0000000	1.9263682
FixedAcidity	0	7.0757171	6.9000000	6.3176435
VolatileAcidity	0	0.3241039	0.2800000	0.7840142
CitricAcid	0	0.3084127	0.3100000	0.8620798
ResidualSugar	616	5.4187331	3.9000000	33.7493790
Chlorides	638	0.0548225	0.0460000	0.3184673
FreeSulfurDioxide	647	30.8455713	30.0000000	148.7145577
TotalSulfurDioxide	682	120.7142326	123.0000000	231.9132105
Density	0	0.9942027	0.9944900	0.0265376
pH	395	3.2076282	3.2000000	0.6796871
Sulphates	1210	0.5271118	0.5000000	0.9321293
Alcohol	653	10.4892363	10.4000000	3.7278190
LabelAppeal	0	-0.0090660	0	0.8910892
AcidIndex	0	7.7727237	8.0000000	1.3239264
STARS	3359	2.0417550	2.0000000	0.9025400

The median and mode for all most all variables are similar. The standard deviation for FreesulphurDioxide and Totalsulfurdioxide is very high

B) Correlation

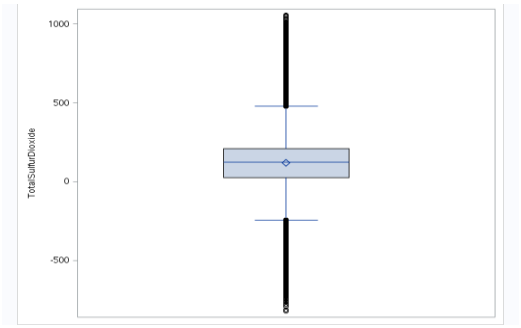
Target and Stars had the highest correlation of 55 % among all the variables. Rest of the variables exhibited no correlation.

TARGET	TARGET 1.00000 12795	STARS 0.55879 <.0001 9436	LabelAppeal 0.35650 <.0001 12795	AcidIndex -0.24605 <.0001 12795	VolatileAcidity -0.08879 <.0001 12795	Alcohol 0.06206 <.0001 12142
FixedAcidity	FixedAcidity 1.00000 12795	AcidIndex 0.17844 <.0001 12795	TARGET -0.04901 <.0001 12795	Sulphates 0.03078 0.0009 11585	TotalSulfurDioxide -0.02250 0.0133 12113	ResidualSugar -0.01885 0.0375 12179
VolatileAcidity	VolatileAcidity 1.00000 12795	TARGET -0.08879 <.0001 12795	AcidIndex 0.04464 <.0001 12795	STARS -0.03443 0.0008 9436	TotalSulfurDioxide -0.02108 0.0203 12113	LabelAppeal -0.01699 0.0547 12795
CitricAcid	CitricAcid 1.00000 12795	AcidIndex 0.06570 <.0001 12795	Alcohol 0.01705 0.0603 12142	VolatileAcidity -0.01695 0.0552 12795	FixedAcidity 0.01424 0.1072 12795	Density -0.01395 0.1145 12795
ResidualSugar	ResidualSugar 1.00000 12179	TotalSulfurDioxide 0.02248 0.0158 11532	Alcohol -0.02000 0.0315 11563	FixedAcidity -0.01885 0.0375 12179	FreeSulfurDioxide 0.01749 0.0600 11563	STARS 0.01674 0.1126 8984
Chlorides	Chlorides 1.00000 12157	TARGET -0.03826 <.0001 12157	AcidIndex 0.02524 0.0054 12157	Density 0.02266 0.0125 12157	FreeSulfurDioxide -0.02066 0.0264 11544	Alcohol -0.01969 0.0344 11538
FreeSulfurDioxide	FreeSulfurDioxide 1.00000 12148	TARGET 0.04382 <.0001 12148	AcidIndex -0.04172 <.0001 12148	Chlorides -0.02066 0.0264 11544	Alcohol -0.01859 0.0460 11527	ResidualSugar 0.01749 0.0600 11563
TotalSulfurDioxide	TotalSulfurDioxide 1.00000 12113	TARGET 0.05148 <.0001 12113	AcidIndex -0.04931 <.0001 12113	FixedAcidity -0.02250 0.0133 12113	ResidualSugar 0.02248 0.0158 11532	VolatileAcidity -0.02108 0.0203 12113
Density	Density 1.00000 12795	AcidIndex 0.04041 <.0001 12795	TARGET -0.03552 <.0001 12795	Chlorides 0.02266 0.0125 12157	STARS -0.01828 0.0757 9436	VolatileAcidity 0.01473 0.0956 12795

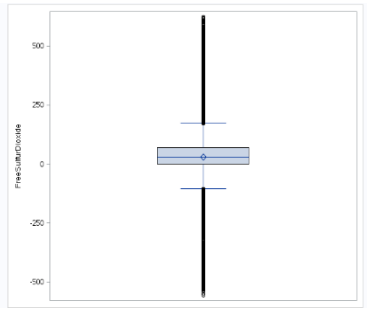
pH	pH 1.00000 12400	AcidIndex -0.05868 <.0001 12400	Chlorides -0.01761 0.0561 11773	VolatileAcidity 0.01359 0.1302 12400	ResidualSugar 0.01212 0.1880 11802
Sulphates	Sulphates 1.00000 11585	TARGET -0.03885 <.0001 11585	AcidIndex 0.03445 0.0002 11585	FixedAcidity 0.03078 0.0009 11585	CitricAcid -0.01299 0.1621 11585
Alcohol	Alcohol 1.00000 12142	STARS 0.06522 <.0001 8963	TARGET 0.06206 <.0001 12142	AcidIndex -0.03814 <.0001 12142	ResidualSugar -0.02000 0.0315 11563
LabelAppeal	LabelAppeal 1.00000 12795	TARGET 0.35650 <.0001 12795	STARS 0.33479 <.0001 9436	AcidIndex 0.02475 0.0051 12795	VolatileAcidity -0.01699 0.0547 12795
AcidIndex	AcidIndex 1.00000 12795	TARGET -0.24605 <.0001 12795	FixedAcidity 0.17844 <.0001 12795	STARS -0.08626 <.0001 9436	CitricAcid 0.06570 <.0001 12795
STARS	STARS 1.00000 9436	TARGET 0.55879 <.0001 9436	LabelAppeal 0.33479 <.0001 9436	AcidIndex -0.08626 <.0001 9436	Alcohol 0.06522 <.0001 8963

C) Box plots of the variables : Outliers
The following two variables exhibited outlier's

a) TotalSulphurdioxide



b) FreeSulphurdioxide



Both the variables were set respectively to their 99% and 1 % respectively to remove the effects of the outlier.

D) The following variables had missing values. The variables had missing values imputed based n decision tree. These can be found in Data Preparation step

The SAS System

The MEANS Procedure

Variable	N Miss	1st Pctl	25th Pctl	50th Pctl	99th Pctl
INDEX	0	166.0000000	4037.00	8110.00	15975.00
TARGET	0	0	2.0000000	3.0000000	7.0000000
FixedAcidity	0	-10.9000000	5.2000000	6.9000000	24.4000000
VolatileAcidity	0	-1.8650000	0.1300000	0.2800000	2.5900000
CitricAcid	0	-2.1800000	0.0300000	0.3100000	2.6600000
ResidualSugar	616	-91.0000000	-2.0000000	3.9000000	99.2000000
Chlorides	638	-0.8590000	-0.0310000	0.0460000	0.9570000
FreeSulfurDioxide	647	-388.0000000	0	30.0000000	469.0000000
TotalSulfurDioxide	682	-531.0000000	27.0000000	123.0000000	767.0000000
Density	0	0.9168000	0.9877200	0.9944900	1.0698100
pH	395	1.3200000	2.9600000	3.2000000	5.1250000
Sulphates	1210	-2.1300000	0.2800000	0.5000000	3.1600000
Alcohol	653	0.1000000	9.0000000	10.4000000	20.3000000
LabelAppeal	0	-2.0000000	-1.0000000	0	2.0000000
AcidIndex	0	6.0000000	7.0000000	8.0000000	13.0000000
STARS	3359	1.0000000	1.0000000	2.0000000	4.0000000
TARGET_FLAG	0	0	1.0000000	1.0000000	1.0000000

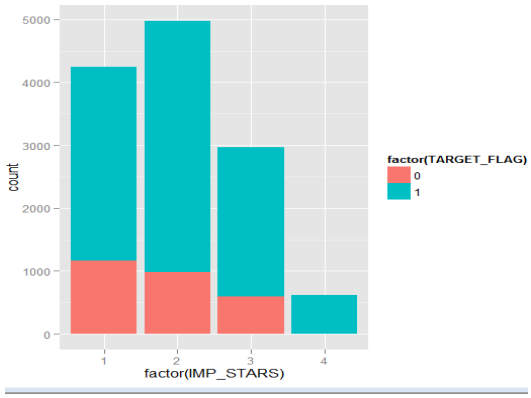
E) Data Distribution

1) Using Random Forest the following is the order of importance in terms of impact

i. STARS
ii. LabelAppeal
iii. Alcohol
iv. Density
v. Chlorides
vi. Volatile Acidity
vii. TotalSulfurDioxide
viii. FreeSulfurDioxide
ix. Citric Acid

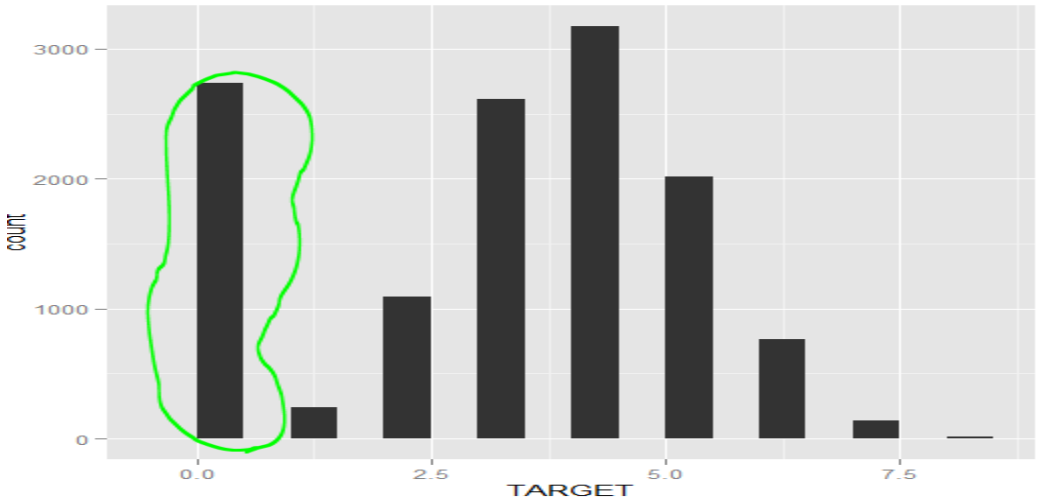
x. Sulphates
xi. pH
xii. ResidualSugar
xiii. Fixed Acidity
xiv. AcidIndex

2) STARS vs TARGET_FLAG



As the STARS value increases, more cases are sold.

3) Target Histogram



Since the number of Zeros are many, Ideally Zero inflated Poisson or Zero inflated Negative binomial distribution should be used.

4) Mean vs Variance of Target

The SAS System

The MEANS Procedure

Analysis Variable : TARGET	
Mean	Variance
3.8522016	1.5482330

As mean > variance, ideally neither Poisson nor Negative Binomial Distribution is suited. Rather, Binomial distribution should be used to count the occurrence .But for the purpose of the assignment Poisson and Negative Binomial would be used.

2. DATA PREPARATION

1) The missing values for these variables were imputed using a decision tree.
   A) ResidualSugar
   B) Chlorides
   C) FreeSulfurDioxide
   D) TotalSulfurDioxide
   E) pH
   F) Sulphates
   G) Alcohol
   H) STARS

Missing value imputation & Flag Creation
Decision Trees (TREE package in R & Weka)

STARS

```
IF MISSING (STARS) THEN DO;
 M_STARS=1;
 IF LabelAppeal >-0.5 AND LabelAppeal < .5 THEN IMP_STARS = 2.00;
 IF LabelAppeal < -0.5 THEN IMP_STARS = 1;
 IF LabelAppeal > .5   THEN IMP_STARS = 3;
 END;
```

Alcohol

```
 IMP_Alcohol = Alcohol;
 IF MISSING (Alcohol) THEN DO;
 IF Density <= .98 THEN IMP_Alcohol = 10.37;
 IF Density > .98   AND DENSITY <=.99 THEN IMP_Alcohol = 11.06;
```

```
IF Density > .99   AND DENSITY <=1.01 THEN IMP_Alcohol = 10.05;
IF Density > 1.01 THEN IMP_Alcohol = 10.42;
END;
```

Chlorides

```
IMP_Chlorides = Chlorides;
IF MISSING (Chlorides) THEN DO;
IF AcidIndex >= 4       AND AcidIndex < 7   THEN IMP_Chlorides = .05;
IF AcidIndex > = 7      AND AcidIndex < 9   THEN IMP_Chlorides = .06;
IF AcidIndex > = 9      THEN IMP_Chlorides = .08;
END;
```

FreeSulphurDioxide

```
IMP_FreeSulfurDioxide = FreeSulfurDioxide;
IF MISSING (FreeSulfurDioxide) THEN DO;
IF    AcidIndex <=9 THEN IMP_FreeSulfurDioxide = 32.83;
    ELSE
    IMP_FreeSulfurDioxide = 9.27     ;
END;
```

Ph

```
IMP_pH = pH;
IF   MISSING (pH) THEN DO;
IF   AcidIndex   <= 6     THEN IMP_pH = 3.32  ;
IF   AcidIndex    =7      THEN  IMP_pH = 3.23;
IF   AcidIndex    >=8     Then IMP_pH = 3.17;
END;
```

TotalSulfurDioxide

```
IMP_TotalSulfurDioxide = TotalSulfurDioxide;
IF MISSING (TotalSulfurDioxide) THEN DO;
IF MISSING(ResidualSugar) THEN IMP_ResidualSugar = 3.9;
IF IMP_ResidualSugar < 1.5   THEN IMP_TotalSulfurDioxide = 112;
IF IMP_ResidualSugar >= 1.5 AND   IMP_ResidualSugar < 6.1 THEN IMP_TotalSulfurDioxide = 100;
IF IMP_ResidualSugar >= 6.1 AND   IMP_ResidualSugar < 19.7 THEN IMP_TotalSulfurDioxide = 148;
IF IMP_ResidualSugar >= 19.7 THEN IMP_TotalSulfurDioxide = 125.89;
END;
```

Sulphates

```
IMP_Sulphates = Sulphates;
IF MISSING (Sulphates) THEN DO;
IF IMP_TotalSulfurDioxide <= 13 THEN IMP_Sulphates = .55;
IF IMP_TotalSulfurDioxide > 13 AND   IMP_TotalSulfurDioxide <= 75 THEN IMP_Sulphates= .63;
IF IMP_TotalSulfurDioxide > 75 AND   IMP_TotalSulfurDioxide <= 169 THEN IMP_Sulphates = .47;
IF IMP_TotalSulfurDioxide > 169 THEN IMP_Sulphates = .53;
END;
```

Residual Sugar

```
IMP_ResidualSugar = ResidualSugar;
IF   MISSING (ResidualSugar) THEN DO;
IF IMP_TotalSulfurDioxide   <=-166 THEN IMP_ResidualSugar = 6.01;
IF IMP_TotalSulfurDioxide   > -166 AND IMP_TotalSulfurDioxide <=137 THEN IMP_ResidualSugar =3.7;
IF IMP_TotalSulfurDioxide   > 137   THEN IMP_ResidualSugar = 7.11;
```

2) Dummy Variable coding.

LabelAppeal and STARS were recorded in Dummy variable format as the following

STARS

a. M_STARS1
b. M_STARS2
c. M_STARS3
d. M_STARS4

LabelAppeal

a. M_LabelAppeal1
b. M_LabelAppeal2
c. M_LabelAppeal3
d. M_LabelAppeal4
e. M_LabelAppeal0

3) Interaction Variables

Interaction between LabelAppeal and STARS:

INT_LABEL_STARS1A = M_STARS1 * M_LabelAppeal1;
INT_LABEL_STARS1B = M_STARS1 * M_LabelAppeal2;
INT_LABEL_STARS1C = M_STARS1 * M_LabelAppeal3;
INT_LABEL_STARS1D = M_STARS1 * M_LabelAppeal4;
INT_LABEL_STARS1E = M_STARS1 * M_LabelAppeal0;

INT_LABEL_STARS2A = M_STARS2 * M_LabelAppeal1;
INT_LABEL_STARS2B = M_STARS2 * M_LabelAppeal2;
INT_LABEL_STARS2C = M_STARS2 * M_LabelAppeal3;
INT_LABEL_STARS2D = M_STARS2 * M_LabelAppeal4;
INT_LABEL_STARS2E = M_STARS2 * M_LabelAppeal0;

INT_LABEL_STARS3A = M_STARS3 * M_LabelAppeal1;
INT_LABEL_STARS3B = M_STARS3 * M_LabelAppeal2;
INT_LABEL_STARS3C = M_STARS3 * M_LabelAppeal3;
INT_LABEL_STARS3D = M_STARS3 * M_LabelAppeal4;
INT_LABEL_STARS3E = M_STARS3 * M_LabelAppeal0;

INT_LABEL_STARS4A = M_STARS4 * M_LabelAppeal1;
INT_LABEL_STARS4B = M_STARS4 * M_LabelAppeal2;
INT_LABEL_STARS4C = M_STARS4 * M_LabelAppeal3;
INT_LABEL_STARS4D = M_STARS4 * M_LabelAppeal4;
INT_LABEL_STARS4E = M_STARS4 * M_LabelAppeal0;

3. BUILD MODELS

Build at least five different using the SAS procs: PROC GENMOD and PROC REG. The five models will be:

1) GENMOD with Poisson distribution

Using a series of models, Based on the lowest AIC the following model was used for Poisson Regression

AIC

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	13495.6231	1.0568
Scaled Deviance	13E3	13495.6231	1.0568
Pearson Chi-Square	13E3	11424.9985	0.8947
Scaled Pearson X2	13E3	11424.9985	0.8947
Log Likelihood		8878.3489	
Full Log Likelihood		-22718.8224	
AIC (smaller is better)		45487.6448	
AICC (smaller is better)		45487.7466	
BIC (smaller is better)		45674.0650	

Parameter Values

Parameter	DF	Estimate	Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.1814	0.0573	1.0691	1.2936	425.26	<.0001
VolatileAcidity	1	-0.0298	0.0065	-0.0426	-0.0170	20.76	<.0001
AcidIndex	1	-0.0775	0.0045	-0.0863	-0.0686	294.99	<.0001
IMP_Chlorides	1	-0.0369	0.0165	-0.0691	-0.0047	5.03	0.0249
IMP_FreeSulfurDioxid	1	0.0001	0.0000	0.0000	0.0002	7.03	0.0080
IMP_TotalSulfurDioxi	1	0.0001	0.0000	0.0000	0.0001	10.58	0.0011
IMP_Sulphates	1	-0.0133	0.0057	-0.0246	-0.0021	5.39	0.0203
IMP_Alcohol	1	0.0039	0.0014	0.0012	0.0067	7.84	0.0051
M_STARS	1	-0.4199	0.0454	-0.5089	-0.3308	85.39	<.0001
INT_LABEL_STARS1A	1	0.5015	0.0505	0.4026	0.6004	98.74	<.0001
INT_LABEL_STARS1B	1	0.3330	0.1008	0.1354	0.5305	10.92	0.0010
INT_LABEL_STARS1C	1	0.2035	0.0474	0.1107	0.2963	18.46	<.0001
INT_LABEL_STARS1E	1	0.4072	0.0456	0.3178	0.4965	79.76	<.0001
INT_LABEL_STARS2A	1	0.8963	0.0454	0.8073	0.9852	389.95	<.0001
INT_LABEL_STARS2B	1	1.0431	0.0593	0.9269	1.1594	309.20	<.0001
INT_LABEL_STARS2C	1	0.4398	0.0470	0.3476	0.5320	87.43	<.0001
INT_LABEL_STARS2E	1	0.6981	0.0444	0.6111	0.7850	247.71	<.0001
INT_LABEL_STARS3A	1	0.9713	0.0455	0.8822	1.0605	456.09	<.0001
INT_LABEL_STARS3B	1	1.1448	0.0540	1.0389	1.2507	449.16	<.0001
INT_LABEL_STARS3C	1	0.6096	0.0538	0.5042	0.7150	128.50	<.0001
INT_LABEL_STARS3E	1	0.8099	0.0452	0.7214	0.8985	321.41	<.0001
INT_LABEL_STARS4A	1	1.0715	0.0489	0.9757	1.1673	480.56	<.0001
INT_LABEL_STARS4B	1	1.2188	0.0612	1.0988	1.3388	396.17	<.0001
INT_LABEL_STARS4C	1	0.8229	0.0978	0.6311	1.0147	70.73	<.0001
INT_LABEL_STARS4E	1	0.9457	0.0536	0.8407	1.0507	311.61	<.0001

2) GENMOD with Negative Binomial distribution

AIC

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	13640.5770	1.0674
Scaled Deviance	13E3	13640.5770	1.0674
Pearson Chi-Square	13E3	11271.3980	0.8820
Scaled Pearson X2	13E3	11271.3980	0.8820
Log Likelihood		8805.8719	
Full Log Likelihood		-22791.2993	
AIC (smaller is better)		45616.5987	
AICC (smaller is better)		45616.6466	
BIC (smaller is better)		45743.3645	

Parameters

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	2.1396	0.0423	2.0566	2.2226	2554.21	<.0001
VolatileAcidity	1	-0.0308	0.0065	-0.0436	-0.0180	22.25	<.0001
AcidIndex	1	-0.0792	0.0045	-0.0880	-0.0703	309.03	<.0001
IMP_Chlorides	1	-0.0385	0.0165	-0.0708	-0.0063	5.48	0.0193
IMP_FreeSulfurDioxid	1	0.0001	0.0000	0.0000	0.0002	7.54	0.0060
IMP_TotalSulfurDioxi	1	0.0001	0.0000	0.0000	0.0001	11.74	0.0006
IMP_Sulphates	1	-0.0128	0.0057	-0.0241	-0.0016	4.98	0.0256
IMP_Alcohol	1	0.0039	0.0014	0.0011	0.0066	7.50	0.0062
M_STARS	1	-1.3256	0.0243	-1.3732	-1.2779	2974.31	<.0001
M_STARS1	1	-0.5592	0.0217	-0.6016	-0.5167	666.04	<.0001
M_STARS2	1	-0.2393	0.0199	-0.2783	-0.2002	144.49	<.0001
M_STARS3	1	-0.1198	0.0202	-0.1594	-0.0802	35.17	<.0001
M_STARS4	0	0.0000	0.0000	0.0000	0.0000	.	.
M_LabelAppeal1	1	0.1323	0.0123	0.1083	0.1563	116.64	<.0001
M_LabelAppeal2	1	0.2694	0.0228	0.2246	0.3142	139.05	<.0001
M_LabelAppeal3	1	-0.1900	0.0143	-0.2180	-0.1619	176.31	<.0001
M_LabelAppeal4	1	-0.4258	0.0371	-0.4984	-0.3532	132.06	<.0001
M_LabelAppeal0	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion	0	0.0000	0.0000	0.0000	0.0000		

3) GENMOD with Zero Inflated Poisson distribution

AIC

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		41016.1246	
Scaled Deviance		41016.1246	
Pearson Chi-Square	13E3	5973.6601	0.4680
Scaled Pearson X2	13E3	5973.6601	0.4680
Log Likelihood		11089.1090	
Full Log Likelihood		-20508.0623	
AIC (smaller is better)		41080.1246	
AICC (smaller is better)		41080.2901	
BIC (smaller is better)		41318.7425	

Parameters

A) Poisson

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.0968	0.0656	-0.0318	0.2254	2.18	0.1402
VolatileAcidity	1	-0.0165	0.0067	-0.0296	-0.0033	6.02	0.0141
AcidIndex	1	-0.0229	0.0048	-0.0324	-0.0135	22.64	<.0001
IMP_Alcohol	1	0.0069	0.0014	0.0041	0.0097	23.02	<.0001
M_STARS	1	0.3050	0.0322	0.2419	0.3681	89.79	<.0001
IMP_STARS	1	0.4251	0.0140	0.3976	0.4526	916.66	<.0001
INT_LABEL_STARS1A	1	1.0331	0.0473	0.9404	1.1258	477.03	<.0001
INT_LABEL_STARS1B	1	1.2209	0.1011	1.0227	1.4191	145.76	<.0001
INT_LABEL_STARS1C	1	0.4259	0.0436	0.3405	0.5114	95.55	<.0001
INT_LABEL_STARS1E	1	0.7680	0.0425	0.6848	0.8512	327.15	<.0001
INT_LABEL_STARS2A	1	0.7038	0.0304	0.6443	0.7633	537.40	<.0001
INT_LABEL_STARS2B	1	0.8733	0.0489	0.7774	0.9691	318.78	<.0001
INT_LABEL_STARS2C	1	0.2327	0.0327	0.1685	0.2968	50.54	<.0001
INT_LABEL_STARS2E	1	0.4992	0.0288	0.4428	0.5557	300.83	<.0001
INT_LABEL_STARS3A	1	0.3461	0.0229	0.3012	0.3911	227.73	<.0001
INT_LABEL_STARS3B	1	0.5119	0.0371	0.4392	0.5846	190.44	<.0001
INT_LABEL_STARS3E	1	0.1810	0.0223	0.1372	0.2248	65.60	<.0001
INT_LABEL_STARS4C	1	-0.2165	0.0895	-0.3920	-0.0410	5.85	0.0156
Scale	0	1.0000	0.0000	1.0000	1.0000		

B) Logit

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.0182	0.3389	-5.6825	-4.3539	219.20	<.0001
VolatileAcidity	1	0.1813	0.0453	0.0925	0.2700	16.03	<.0001
AcidIndex	1	0.4277	0.0263	0.3762	0.4793	264.38	<.0001
IMP_FreeSulfurDioxid	1	-0.0009	0.0003	-0.0014	-0.0003	10.54	0.0012
IMP_TotalSulfurDioxi	1	-0.0011	0.0002	-0.0014	-0.0007	40.30	<.0001
IMP_pH	1	0.2181	0.0530	0.1143	0.3220	16.95	<.0001
IMP_Sulphates	1	0.1399	0.0396	0.0622	0.2176	12.46	0.0004
IMP_Alcohol	1	0.0284	0.0099	0.0089	0.0478	8.13	0.0044
M_STARS	1	5.1662	0.1642	4.8444	5.4879	990.34	<.0001
IMP_STARS	1	-1.9117	0.1041	-2.1158	-1.7077	337.22	<.0001
M_LabelAppeal1	1	1.8237	0.1212	1.5862	2.0612	226.48	<.0001
M_LabelAppeal2	1	2.1245	0.1964	1.7395	2.5095	116.98	<.0001
M_LabelAppeal3	1	-2.9793	0.1552	-3.2834	-2.6752	368.69	<.0001
M_LabelAppeal4	1	-3.8840	0.2597	-4.3930	-3.3750	223.71	<.0001
M_LabelAppeal0	0	0.0000	0.0000	0.0000	0.0000	.	.

4) GENMOD with Zero Inflated Negative Binomial distribution

AIC

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		41010.6560	
Scaled Deviance		41010.6560	
Pearson Chi-Square	13E3	5922.6736	0.4638
Scaled Pearson X2	13E3	5922.6736	0.4638
Log Likelihood		-20505.3280	
Full Log Likelihood		-20505.3280	
AIC (smaller is better)		41062.6560	
AICC (smaller is better)		41062.7660	
BIC (smaller is better)		41256.5331	

Parameters

A) Negative Binomial

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.6543	0.0445	1.5671	1.7415	1382.55	<.0001
AcidIndex	1	-0.0244	0.0048	-0.0338	-0.0149	25.59	<.0001
IMP_Alcohol	1	0.0069	0.0014	0.0040	0.0097	22.88	<.0001
M_STARS	1	-0.3704	0.0256	-0.4205	-0.3202	209.73	<.0001
M_STARS1	1	-0.3527	0.0223	-0.3964	-0.3090	250.68	<.0001
M_STARS2	1	-0.1885	0.0199	-0.2275	-0.1494	89.39	<.0001
M_STARS3	1	-0.0972	0.0202	-0.1368	-0.0576	23.17	<.0001
M_STARS4	0	0.0000	0.0000	0.0000	0.0000	.	.
M_LabelAppeal1	1	0.1950	0.0125	0.1704	0.2195	242.10	<.0001
M_LabelAppeal2	1	0.3513	0.0231	0.3060	0.3966	231.04	<.0001
M_LabelAppeal3	1	-0.3046	0.0149	-0.3338	-0.2755	419.17	<.0001
M_LabelAppeal4	1	-0.7221	0.0397	-0.7999	-0.6443	331.23	<.0001
Dispersion	0	0.0000	0.0000	0.0000	0.0000		

B) Logit

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.9962	0.3366	-5.6559	-4.3366	220.37	<.0001
VolatileAcidity	1	0.1917	0.0449	0.1037	0.2797	18.24	<.0001
AcidIndex	1	0.4264	0.0261	0.3752	0.4776	266.04	<.0001
IMP_FreeSulfurDioxid	1	-0.0009	0.0003	-0.0014	-0.0003	10.22	0.0014
IMP_TotalSulfurDioxi	1	-0.0011	0.0002	-0.0014	-0.0008	40.84	<.0001
IMP_pH	1	0.2180	0.0527	0.1147	0.3213	17.12	<.0001
IMP_Sulphates	1	0.1417	0.0394	0.0645	0.2188	12.96	0.0003
IMP_Alcohol	1	0.0279	0.0099	0.0086	0.0473	8.01	0.0047
M_STARS	1	5.1039	0.1610	4.7883	5.4194	1005.04	<.0001
IMP_STARS	1	-1.8859	0.1034	-2.0885	-1.6833	332.97	<.0001
M_LabelAppeal1	1	1.7687	0.1191	1.5354	2.0021	220.66	<.0001
M_LabelAppeal2	1	2.0766	0.1948	1.6949	2.4583	113.70	<.0001
M_LabelAppeal3	1	-2.8454	0.1527	-3.1448	-2.5460	347.05	<.0001
M_LabelAppeal4	1	-4.3453	0.3481	-5.0276	-3.6630	155.81	<.0001

5) REGRESSION (use standard PROC REG and Stepwise selection)

AIC & Adjusted R square

The SAS System

Obs	_MODEL_	_AIC_	_SBC_	_BIC_	_CP_	_ADJRSQ_
1	model2	6189.67	6391.00	6191.78	27	0.56379

Annova Table

The SAS System

The REG Procedure
Model: model2
Dependent Variable: TARGET

Number of Observations Read	12795
Number of Observations Used	12795

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	26	26809	1031.11899	636.99	<.0001
Error	12768	20668	1.61874		
Corrected Total	12794	47477			

Root MSE	1.27230	R-Square	0.5647
Dependent Mean	3.02907	Adj R-Sq	0.5638
Coeff Var	42.00286		

Parameters with VIF in last column

Intercept	1	4.16213	0.43669	9.53	<.0001	0
VolatileAcidity	1	-0.08815	0.01440	-6.12	<.0001	1.00765
Density	1	-0.87120	0.42474	-2.05	0.0403	1.00415
AcidIndex	1	-0.18832	0.00871	-21.63	<.0001	1.05017
IMP_Chlorides	1	-0.10990	0.03629	-3.03	0.0025	1.00323
IMP_FreeSulfurDioxide	1	0.00029785	0.00008488	3.51	0.0005	1.00520
IMP_TotalSulfurDioxide	1	0.00022071	0.00005404	4.08	<.0001	1.00539
IMP_pH	1	-0.03296	0.01686	-1.96	0.0505	1.00553
IMP_Sulphates	1	-0.03778	0.01270	-2.98	0.0029	1.00272
IMP_Alcohol	1	0.01325	0.00311	4.26	<.0001	1.00807
M_STARS	1	-0.57459	0.07756	-7.41	<.0001	9.20541
INT_LABEL_STARS1A	1	1.21780	0.09557	12.74	<.0001	2.43925
INT_LABEL_STARS1B	1	0.74555	0.19649	3.79	0.0001	1.16426
INT_LABEL_STARS1C	1	0.43172	0.08437	5.12	<.0001	4.08320
INT_LABEL_STARS1E	1	0.94819	0.08203	11.56	<.0001	4.96717
INT_LABEL_STARS2A	1	2.71857	0.08585	31.67	<.0001	3.70373
INT_LABEL_STARS2B	1	3.48063	0.14271	24.39	<.0001	1.35972
INT_LABEL_STARS2C	1	1.04239	0.08612	12.10	<.0001	3.63208
INT_LABEL_STARS2E	1	1.90633	0.08051	23.68	<.0001	5.81086
INT_LABEL_STARS3A	1	3.12206	0.08732	35.75	<.0001	3.39225
INT_LABEL_STARS3B	1	4.04776	0.12713	31.84	<.0001	1.49954
INT_LABEL_STARS3C	1	1.59291	0.10814	14.73	<.0001	1.85410
INT_LABEL_STARS3E	1	2.36139	0.08432	28.01	<.0001	4.08941
INT_LABEL_STARS4A	1	3.65931	0.10364	35.31	<.0001	2.00699
INT_LABEL_STARS4B	1	4.48326	0.15970	28.07	<.0001	1.26813
INT_LABEL_STARS4C	1	2.46733	0.24776	9.96	<.0001	1.09725
INT_LABEL_STARS4E	1	2.99652	0.11810	25.37	<.0001	1.62958

6) GENMOD with Logistic Hurdle

1) Logit AIC

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	13277.788	7907.106
SC	13285.245	7989.131
-2 Log L	13275.788	7885.106

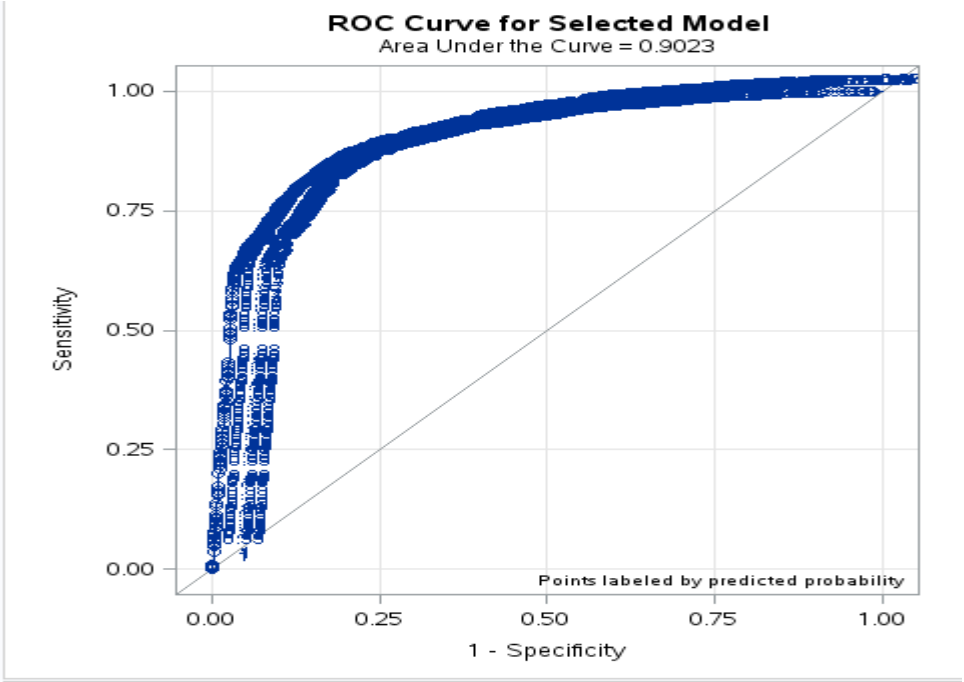
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	5390.6826	10	<.0001
Score	5102.5587	10	<.0001
Wald	2536.2196	10	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
459.3242	20	<.0001

Parameters

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	8.3535	0.2450	1162.1971	<.0001
VolatileAcidity	1	-0.1821	0.0361	25.4779	<.0001
AcidIndex	1	-0.3827	0.0208	338.0770	<.0001
IMP_TotalSulfurDioxi	1	0.000962	0.000137	49.5354	<.0001
M_STARS	1	-4.9588	0.1261	1546.2664	<.0001
IMP_STARS	1	-0.4407	0.0481	84.0688	<.0001
INT_LABEL_STARS1A	1	-3.9923	0.1766	510.9212	<.0001
INT_LABEL_STARS1B	1	-4.7546	0.3284	209.6155	<.0001
INT_LABEL_STARS1C	1	-3.1860	0.1669	364.5261	<.0001
INT_LABEL_STARS1E	1	-3.5444	0.1564	513.3577	<.0001
INT_LABEL_STARS2C	1	-1.2517	0.2161	33.5379	<.0001

ROC



2) Poisson Regression

AIC

Number of Observations Read	12795
Number of Observations Used	10061
Missing Values	2734

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	1E4	2830.1077	0.2816
Scaled Deviance	1E4	2830.1077	0.2816
Pearson Chi-Square	1E4	2578.1335	0.2565
Scaled Pearson X2	1E4	2578.1335	0.2565
Log Likelihood		3061.6642	
Full Log Likelihood		-15582.7759	
AIC (smaller is better)		31187.5519	
AICC (smaller is better)		31187.5782	
BIC (smaller is better)		31266.9325	

Parameters

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.3595	0.0503	1.2610	1.4580	731.74	<.0001
AcidIndex	1	-0.0207	0.0054	-0.0313	-0.0100	14.44	0.0001
IMP_Alcohol	1	0.0092	0.0016	0.0060	0.0124	32.06	<.0001
M_STARS	1	-0.4373	0.0285	-0.4931	-0.3815	236.01	<.0001
M_STARS1	1	-0.3753	0.0247	-0.4237	-0.3269	230.75	<.0001
M_STARS2	1	-0.2342	0.0222	-0.2778	-0.1906	110.79	<.0001
M_STARS3	1	-0.1244	0.0225	-0.1684	-0.0803	30.60	<.0001
M_STARS4	0	0.0000	0.0000	0.0000	0.0000	.	.
M_LabelAppeal1	1	0.2410	0.0140	0.2135	0.2684	295.48	<.0001
M_LabelAppeal2	1	0.4331	0.0254	0.3833	0.4829	290.65	<.0001
M_LabelAppeal3	1	-0.3745	0.0175	-0.4087	-0.3402	459.18	<.0001
M_LabelAppeal4	1	-1.0265	0.0531	-1.1305	-0.9225	374.22	<.0001
M_LabelAppeal0	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale	0	1.0000	0.0000	1.0000	1.0000		

Prasanna Krishna Rao

Parameter and Coefficient Explanation

	Poisson Distribu	Negative Binomial	ZIP POI	ZINB NB	Regression	Logistic Hurdle
AcidIndex	Negative	Negative Impact	Positive	Negative	Negative	Negative
Alcohol	Positive Impact	Positive Impact	Positive	Positive Impact	Positive Impact	Positive Impact
Chlorides	Negative	Negative Impact	Not Significant	Not Significant	Negative	Not Significant
CitricAcid	Not Significant	Not Significant	Not Significant	Not Significant	Not Significant	Not Significant
Density	Not Significant	Not Significant	Not Significant	Not Significant	Negative	Not Significant
FixedAcidity	Not Significant	Not Significant	Not Significant	Not Significant	Not Significant	Not Significant
FreeSulfurDioxi	Positive Impact	Positive Impact	Not Significant	Not Significant	Positive Impact	Not Significant
LabelAppeal	Positive Impact	Positive Impact	Positive	Positive Impact	Positive Impact	
ResidualSugar	Not Significant	Not Significant	Not Significant	Not Significant	Not Significant	Not Significant
STARS	Positive Impact	Positive Impact	Positive Impact	Positive Impact	Positive Impact	
Sulphates	Negative Impact	Negative Impact	Not Significant	Not Significant	Negative Impact	Not Significant
TotalSulfurDioxide	Positive Impact	Positive Impact	Not Significant	Not Significant	Positive Impact	Not Significant
VolatileAcidity	Negative Impact	Negative Impact	Negative Impact	Not Significant	Negative Impact	Not Significant
pH	Not Significant	Not Significant	Not Significant	Not Significant	Positive Impact	Not Significant
Stars Missing Flag	Negative Impact	Negative Impact	Negative Impact	Negative Impact	Negative Impact	Negative Impact

1) Acid Index has a negative impact for all models except for ZIP POI model.
2) Alcohol has a positive impact for all models.
3) Chlorides was not statistically significant in ZIP ,ZINB and Logistic Hurdle models Whereas had a negative impact on rest of the models
4) Citric Acid was found to be statistically non-significant in all models.
5) Density : Except for Regression model , Density was not significant in rest of the models
6) Fixed Acidity was found to be statistically non-significant in all models
7) FreeSulphurDioxide & TotalSulphurDioxide. It has a positive impact for Regression, POI, NB models and no impact for rest of the models.
8) Residual Sugars has no statistical significance on any of the models.
9) Sulphates has no positive impact on any of the models. It has a negative impact on POI,NB and regression models
10) VolaitileAcidity has no positive impact on any of the models. It has a negative impact on POI,NB ZIP and regression models
11) pH has no impact on any of the models except for Regression models wherein it has a positive impact.
12) Stars flag missing has a negative impact on all models.

Stars and Label Appeal Interactions

---

1) Maximum positive impact when LabelAppeal is 2 and stars rating was 4 for both Poisson and Negative binomial Regression
2) Maximum positive impact happened when label appeal was 4 and stars rating was 2 for ZIP
3) Maximum positive impact when LabelAppeal is 2 and stars rating was 4 for ZINB

4. SELECT MODELS

Each of the model was selected based on lowest AIC values within the same kind of models .However AIC, Adjusted R square, BIC cannot be used to compare regression with POI or NB or ZIP or Logistic Hurdle.

The accuracy of the various models were tested by taking a simple average or "mean" value and comparing that with the respective scores of the various models. The model with minimum of average error and mean square error was chosen as the chosen model.

The SAS System

Obs	P	_FREQ_	ERROR_MEAN	ERROR_REG	ERROR_POI	ERROR_NB	ERROR_ZIP	ERROR_ZINB	ERROR_HURDLE	ERROR_ENSEMBLE
1	1	12795	1.55924	0.98012	1.00764	1.02451	0.97752	0.97899	0.96374	0.97883

The SAS System

Obs	P	_FREQ_	ERROR_MEAN	ERROR_REG	ERROR_POI	ERROR_NB	ERROR_ZIP	ERROR_ZINB	ERROR_HURDLE	ERROR_ENSEMBLE
1	1.5	12795	1.76929	1.12944	1.15358	1.16731	1.13516	1.13688	1.12002	1.12993

The SAS System

Obs	P	_FREQ_	ERROR_MEAN	ERROR_REG	ERROR_POI	ERROR_NB	ERROR_ZIP	ERROR_ZINB	ERROR_HURDLE	ERROR_ENSEMBLE
1	2	12795	1.92629	1.27095	1.29212	1.30108	1.28381	1.28556	1.26677	1.27108

Based on the Average Error, Constant error and Mean Error, Logistic Hurdle Model was chosen to be the more accurate than rest of the models.

Prasanna Krishna Rao

STAND ALONE SCORING PROGRAM

DEPLOYMENT CODE

```
libname p411 '/folders/myfolders/411' ;
%let INFILE      =  p411.wine_test ;
%let TEMPFILE              =  p411.TEMPFILE;
%let TEMPFILE1             =  p411.TEMPFILE1;
%let TEMPFILE2             =  p411.TEMPFILE2;
%let SCOREFILE             =  p411.Ensemble;



DATA &tempfile.;
    SET &INFILE. ;
    TARGET_FLAG = (TARGET > 0) ;
    TARGET_AMT = TARGET - 1;
    IF TARGET_FLAG = 0 THEN TARGET_AMT = . ;
    run;


 DATA &tempfile1.;
     SET &tempfile.;

   IMP_STARS = STARS;
   M_STARS=0;
   IF MISSING(STARS) THEN DO ;
   M_STARS=1;
   IF  LabelAppeal >-0.5  AND LabelAppeal < .5 THEN  IMP_STARS = 2.00;
   IF  LabelAppeal < -0.5 THEN IMP_STARS = 1;
   IF  LabelAppeal > .5   THEN IMP_STARS = 3;
   END;

   IMP_Alcohol = Alcohol;
   IF MISSING(Alcohol) THEN DO ;
   IF  Density <= .98  THEN  IMP_Alcohol = 10.37  ;
   IF  Density > .98   AND DENSITY <=.99  THEN  IMP_Alcohol = 11.06  ;
   IF  Density > .99   AND DENSITY <=1.01 THEN  IMP_Alcohol = 10.05  ;
   IF  Density > 1.01  THEN  IMP_Alcohol = 10.42  ;
   END;

   IMP_Chlorides = Chlorides;
   IF MISSING(Chlorides) THEN DO ;
   IF  AcidIndex  >= 4       AND  AcidIndex < 7   THEN  IMP_Chlorides = .05  ;
   IF  AcidIndex  > = 7      AND  AcidIndex < 9   THEN  IMP_Chlorides = .06  ;
   IF  AcidIndex  > = 9      THEN IMP_Chlorides  = .08 ;
   END;

   IMP_FreeSulfurDioxide = FreeSulfurDioxide;
   IF MISSING(FreeSulfurDioxide) THEN DO ;
   IF    AcidIndex <=9 THEN IMP_FreeSulfurDioxide = 32.83 ;
       ELSE
       IMP_FreeSulfurDioxide = 9.27    ;
   END;

   IMP_pH = pH;
   IF   MISSING(pH) THEN DO ;
   IF   AcidIndex   <= 6     THEN  IMP_pH = 3.32  ;
   IF   AcidIndex   =7      THEN  IMP_pH = 3.23 ;
   IF   AcidIndex   >=8     Then  IMP_pH = 3.17 ;
   END;

   IMP_TotalSulfurDioxide = TotalSulfurDioxide;
   IF MISSING(TotalSulfurDioxide) THEN DO ;
   IF MISSING(ResidualSugar) THEN IMP_ResidualSugar = 3.9;
   IF  IMP_ResidualSugar < 1.5   THEN  IMP_TotalSulfurDioxide = 112 ;
   IF  IMP_ResidualSugar >= 1.5  AND   IMP_ResidualSugar  < 6.1  THEN IMP_TotalSulfurDioxide = 100 ;
   IF  IMP_ResidualSugar >= 6.1  AND   IMP_ResidualSugar  < 19.7 THEN IMP_TotalSulfurDioxide = 148 ;
   IF  IMP_ResidualSugar >= 19.7 THEN  IMP_TotalSulfurDioxide = 125.89;
   END;

   IMP_Sulphates = Sulphates ;
   IF  MISSING(Sulphates) THEN DO ;
```

```
 IF  IMP_TotalSulfurDioxide <= 13 THEN  IMP_Sulphates = .55  ;
 IF  IMP_TotalSulfurDioxide > 13  AND   IMP_TotalSulfurDioxide  <= 75  THEN IMP_Sulphates= .63 ;
 IF  IMP_TotalSulfurDioxide > 75  AND   IMP_TotalSulfurDioxide  <= 169 THEN IMP_Sulphates = .47 ;
 IF  IMP_TotalSulfurDioxide > 169 THEN  IMP_Sulphates = .53;
 END;

 IMP_ResidualSugar = ResidualSugar;
 IF   MISSING(ResidualSugar) THEN DO ;
 IF IMP_TotalSulfurDioxide   <=-166  THEN IMP_ResidualSugar = 6.01  ;
 IF IMP_TotalSulfurDioxide   > -166  AND  IMP_TotalSulfurDioxide <=137  THEN  IMP_ResidualSugar = 3.77;
 IF IMP_TotalSulfurDioxide   > 137   THEN IMP_ResidualSugar = 7.11;
 END;
 RUN;


DATA &tempfile2.;
  SET &tempfile1.;
   IF FixedAcidity  >=  24  THEN FixedAcidity =24;
   IF FixedAcidity  <=  -9  THEN FixedAcidity =-9;


   IF IMP_FreeSulfurDioxide >= 340 THEN IMP_FreeSulfurDioxide = 340;
   IF IMP_FreeSulfurDioxide <= -300 THEN IMP_FreeSulfurDioxide =-300;

   IF IMP_TotalSulfurDioxide >= 700 THEN IMP_TotalSulfurDioxide = 700;
   IF IMP_TotalSulfurDioxide <= -330 THEN IMP_TotalSulfurDioxide =-330;

    M_STARS1 =          STARS   in ('1');
    M_STARS2 =          STARS   in ('2');
    M_STARS3 =          STARS   in ('3');
    M_STARS4 =          STARS   in ('4');

    M_LabelAppeal1  =  LabelAppeal   in ('1');
    M_LabelAppeal2  =  LabelAppeal   in ('2');
    M_LabelAppeal3  =  LabelAppeal   in ('-1');
    M_LabelAppeal4  =  LabelAppeal   in ('-2');
    M_LabelAppeal0  =  LabelAppeal   in ('0');

    INT_LABEL_STARS1A = M_STARS1 * M_LabelAppeal1 ;
    INT_LABEL_STARS1B = M_STARS1 * M_LabelAppeal2 ;
    INT_LABEL_STARS1C = M_STARS1 * M_LabelAppeal3 ;
    INT_LABEL_STARS1D = M_STARS1 * M_LabelAppeal4 ;
    INT_LABEL_STARS1E = M_STARS1 * M_LabelAppeal0 ;

    INT_LABEL_STARS2A = M_STARS2 * M_LabelAppeal1 ;
    INT_LABEL_STARS2B = M_STARS2 * M_LabelAppeal2 ;
    INT_LABEL_STARS2C = M_STARS2 * M_LabelAppeal3 ;
    INT_LABEL_STARS2D = M_STARS2 * M_LabelAppeal4 ;
    INT_LABEL_STARS2E = M_STARS2 * M_LabelAppeal0 ;

    INT_LABEL_STARS3A = M_STARS3 * M_LabelAppeal1 ;
    INT_LABEL_STARS3B = M_STARS3 * M_LabelAppeal2 ;
    INT_LABEL_STARS3C = M_STARS3 * M_LabelAppeal3 ;
    INT_LABEL_STARS3D = M_STARS3 * M_LabelAppeal4 ;
    INT_LABEL_STARS3E = M_STARS3 * M_LabelAppeal0 ;

    INT_LABEL_STARS4A = M_STARS4 * M_LabelAppeal1 ;
    INT_LABEL_STARS4B = M_STARS4 * M_LabelAppeal2 ;
    INT_LABEL_STARS4C = M_STARS4 * M_LabelAppeal3 ;
    INT_LABEL_STARS4D = M_STARS4 * M_LabelAppeal4 ;
    INT_LABEL_STARS4E = M_STARS4 * M_LabelAppeal0 ;



 run;

DATA &SCOREFILE. ;
    SET &tempfile2. ;

        ******** Regression Model ***************** ;
    Reg_TARGET  =  4.16213 +  VolatileAcidity*(-0.08815) + Density*(-0.87120) + AcidIndex *(-0.18832)
       + IMP_Chlorides*(-0.10990)+ IMP_FreeSulfurDioxide*(0.00029785) + IMP_TotalSulfurDioxide*(0.00022071)
       + IMP_pH*(-0.03296)+ IMP_Sulphates*(-0.03778)+ IMP_Alcohol*(0.01325)+ M_STARS*(-0.57459)
       + INT_LABEL_STARS1A *(1.21780) + INT_LABEL_STARS1B*(0.74555) + INT_LABEL_STARS1C*(0.43172)
       + INT_LABEL_STARS1E*(0.94819)  + INT_LABEL_STARS2A*(2.71857) + INT_LABEL_STARS2B*(3.48063)
```

```
    + INT_LABEL_STARS2C*(1.04239)  + INT_LABEL_STARS2E*(1.90633) + INT_LABEL_STARS3A*(3.12206)
    + INT_LABEL_STARS3B*(4.04776)  + INT_LABEL_STARS3C*(1.59291) + INT_LABEL_STARS3E*(2.36139)
    + INT_LABEL_STARS4A*(3.65931)  + INT_LABEL_STARS4B*(4.48326) + INT_LABEL_STARS4C*(2.46733)
    + INT_LABEL_STARS4E*(2.99652);


   ******** POI Model ****************** ;
  POI_TARGET =  1.1814 + VolatileAcidity*(-0.0298) + AcidIndex*(-0.0775) + IMP_Chlorides*(-0.0369)
   + IMP_FreeSulfurDioxide*(.0001)    + IMP_TotalSulfurDioxide*(.0001) + IMP_Sulphates*(-0.0133)
   + IMP_Alcohol*(0.0039)           + M_STARS*(-0.4199) + INT_LABEL_STARS1A*(0.5015)
   + INT_LABEL_STARS1B*(0.3330) + INT_LABEL_STARS1C*(0.2035) + INT_LABEL_STARS1E*(0.4072)
   + INT_LABEL_STARS2A*(0.8963) + INT_LABEL_STARS2B*(1.0431) + INT_LABEL_STARS2C*(0.4398)
   + INT_LABEL_STARS2E*(0.6981) + INT_LABEL_STARS3A*(0.9713) + INT_LABEL_STARS3B*(1.1448)
   + INT_LABEL_STARS3C*(0.6096) + INT_LABEL_STARS3E*(0.8099) + INT_LABEL_STARS4A*(1.0715)
   + INT_LABEL_STARS4B*(1.2188) + INT_LABEL_STARS4C*(0.8229) + INT_LABEL_STARS4E*(0.9457)
   ;
  POI_TARGET = EXP(POI_TARGET);

  ******** NB Model ****************** ;
  NB_TARGET = 2.1396 + VolatileAcidity*(-0.0308) + AcidIndex *(-0.0792) + IMP_Chlorides*(-0.0385)
     + IMP_FreeSulfurDioxide*(0.0001)        + IMP_TotalSulfurDioxide *(0.0001)
     + IMP_Sulphates*(-0.0128) + IMP_Alcohol*(0.0039)    + M_STARS *(-1.3256)
     + M_STARS1*(-0.5592)     + M_STARS2 *(-0.2393)     + M_STARS3*(-0.1198)
     + M_LabelAppeal1*(0.1323) + M_LabelAppeal2*(0.2694) + M_LabelAppeal3*(-0.1900)
     + M_LabelAppeal4*(-0.425);
     NB_TARGET = EXP(NB_TARGET);


   ******** Logistic Hurdle Model ****************** ;
   *** Logistic Hurdle : Logistic Regression Probability ;
   TEMP = 8.3535 + VolatileAcidity *(-0.1821) + AcidIndex*(-0.3827) + IMP_TotalSulfurDioxide*(0.000962)
        + M_STARS*(-4.9588)+ IMP_STARS*(-0.4407)+ INT_LABEL_STARS1A*(-3.9923)
        + INT_LABEL_STARS1B *(-4.7546)        +  INT_LABEL_STARS1C *(-3.1860) + INT_LABEL_STARS1E*(-3.5444)
        + INT_LABEL_STARS2C*(-1.2517);

   P_TARGET_FLAG = exp(TEMP);
   P_TARGET_FLAG= (P_TARGET_FLAG)/(1+P_TARGET_FLAG)   ;

   Hurdle_Target =  1.3595 +  AcidIndex*(-0.0207) + IMP_Alcohol*(.0092) + M_STARS*(-0.4373)
              + M_STARS1*(-0.3753) + M_STARS2*(-0.2342) + M_STARS3*(-0.1244)
              + M_labelAppeal1*(0.2410)+M_LabelAppeal2*(0.4331)
              + M_LabelAppeal3*(-0.3745)+ M_LabelAppeal4*(-1.0265)     ;
   Hurdle_Target = exp(Hurdle_Target);

   ***Logistic Hurdle : Probability * target_amt  *******        ;
   P_HURDLE_TARGET = P_TARGET_FLAG * (Hurdle_Target+1);


***ZIP POI *****;
 TEMP     = 0.0968 +  VolatileAcidity*(-0.0165)+ AcidIndex*(-0.0229)+IMP_Alcohol*(0.0069)
     + M_STARS*(0.3050) + IMP_STARS*(0.4251) + INT_LABEL_STARS1A*(1.0331)
     + INT_LABEL_STARS1B*(1.2209) + INT_LABEL_STARS1C*(0.4259) + INT_LABEL_STARS1E*(0.7680)
     + INT_LABEL_STARS2A*(0.7038) + INT_LABEL_STARS2B*(0.8733) + INT_LABEL_STARS2C*(0.2327)
     + INT_LABEL_STARS2E*(0.4992) + INT_LABEL_STARS3A*(0.3461) + INT_LABEL_STARS3B*(0.5119)
     + INT_LABEL_STARS3E*(0.1810) + INT_LABEL_STARS4C*(-0.2165)      ;

P_SCORE_ZIP_ALL = exp( TEMP );

 TEMP = -5.0182 + VolatileAcidity*(0.1813) + AcidIndex*(0.4277) + IMP_FreeSulfurDioxide*(-0.0009)
   + IMP_TotalSulfurDioxide*(-0.0011) + IMP_pH*(0.2181) + IMP_Sulphates*(0.1399)
   + IMP_Alcohol*(0.0284) + M_STARS*(5.1662) + IMP_STARS*(-1.9117) + M_LabelAppeal1*(1.8237)
   + M_LabelAppeal2*(2.1245) + M_LabelAppeal3*(-2.9793) + M_LabelAppeal4*(-3.8840);

P_SCORE_ZERO = exp(TEMP)/(1+exp(TEMP));
P_SCORE_ZIP = P_SCORE_ZIP_ALL * (1-P_SCORE_ZERO);

****   ZIP  NB;

TEMP = 1.6543 +  AcidIndex*(-0.0244) + IMP_Alcohol*(0.0069) + M_STARS*(-0.3704)
  +  M_STARS1*(-0.3527)     + M_STARS2*(-0.1885)   + M_STARS3*(-0.0972)
  +  M_LabelAppeal1*(0.1950) + M_LabelAppeal2*(0.3513) + M_LabelAppeal3*(-0.3046)
  +  M_LabelAppeal4*(-0.7221) ;
P_SCORE_ZINB_ALL = exp( TEMP );
```

```
TEMP = -4.9962 + VolatileAcidity*(0.1917) + AcidIndex*(0.4264) + IMP_FreeSulfurDioxide*(-0.0009)
    +  IMP_TotalSulfurDioxide*(-0.0011) + IMP_pH*(0.2180) + IMP_Sulphates*(0.1417) + IMP_Alcohol*(0.0279)
    +  M_STARS*(5.1039) + IMP_STARS*(-1.8859) + M_LabelAppeal1*(1.7687) + M_LabelAppeal2*(2.0766)
    +  M_LabelAppeal3*(-2.8454) + M_LabelAppeal4*(-4.3453) ;

P_SCORE_ZERO = exp(TEMP)/(1+exp(TEMP));
P_SCORE_ZINB = P_SCORE_ZIP_ALL * (1-P_SCORE_ZERO);


 ENSEMBLE = ROUND(( Reg_TARGET + POI_TARGET + NB_TARGET+P_HURDLE_TARGET+P_SCORE_ZIP+P_SCORE_ZINB )/6) ;
 P_TARGET = ENSEMBLE ;
 KEEP INDEX P_TARGET;
 RUN;
PROC PRINT DATA = &SCOREFILE;
```

SCORED DATA FILE

I have included two models. One being the best model and other being ensemble model

scorefile.sas7bdat

1) Model1 :=> Best model using Logistic Hurdle

ensemble.sas7bdat

2) Model 2 :=> Model using Ensemble : Regression, Logistic Hurdle,POI,NB,ZOP,ZNB

Conclusion:

1) As mean is greater than the variance, ideally neither Poisson nor Negative Binomial Distribution is suited. Rather, Binomial distribution should be used to count the occurrence .But for the purpose of the assignment Poisson , Negative Binomial ,ZIP ,ZINB were used
2) Logistic Hurdle Regression model returned minimum mean errors and hence was the model of choice
3) Alcohol , TotalSulphurDioxide,FreeSulphurDioxide,Label Appeal and Stars had maximum impact on the sale of wine
4) Regarding the positive interactions between Label and Stars, A LabelAppeal of 2 and stars rating of 4 had the maximum impact.
5) AcidIndex, Sulphates and Volatile Index had negative impact on most of the models.
6) Citric Acid, Residual Sugars, Chlorides & pHlevel had no impact statistically.