# Assignment #3: Wine Sales Project

PREDICT 411 Section 55

Scott M. Morgan

**KAGGLE:** ScottMorgan

OPUS ONE

DATE: August 20, 2017

TO: Dr. Donald Wedding, Master Sommelier
**President and Vineyard Manager, Opus One**

FROM: Mr. Scott Morgan
**Director of Sales and Vineyard Analytics, Opus One**

SUBJECT: One Vision. One Wine. One Predictive Model.

---

## EXECUTIVE SUMMARY

According to Rob McMillan, founder of Silicon Valley Bank's Wine Division, the dynamics of the wine industry within the United States are at an interesting inflection point. Per capita consumption faces crosscurrents with retiring wine-loyal baby boomers being replaced by less affluent millennials who are indecisive about their alcoholic beverage of choice. If economic conditions continue to improve, however, per capita consumption should be slightly higher in 2018 and beyond. Tangentially, millennials are beginning to affect the lower price range of premium sales, Opus One's key market segment. Their presence is most visible in the $8 to $11.99 red blend category, but research suggests they gradually will shift from blends to varietal wines or imports as their incomes grow. It is imperative that our organization takes a proactive approach to understanding not only our current offerings but preparing for the preferences of a younger customer base.

It has come to the attention of senior leadership the need to devise a robust system of identifying which characteristics of Opus One wine are drivers of sales. At a fundamental level, the factors that increase customers' propensity to purchase our varietals is a source of great interest to our business and stakeholders. The knowledge of the quantitative and qualitative characteristics that appeal to customers, and the development of predictive models based on those features, is important across the company, from the Tasting Room/Wine Club Manager to the Cellar Master.

Overall, this exercise is intended to ensure Opus One is utilizing forward-looking analysis in our sales and product development efforts. The following report is a detailed account of the predictive modeling process that accomplishes this.

To summarize, the ***key findings*** are that missing ratings, label appeal and the presence of the ratings themselves tend to be highly predictive of how much wine is purchased. The ***basic managerial recommendation*** is to focus marketing efforts on organizations which provide tasting reviews (Wine Spectator, Wine Enthusiast, etc.) to ensure our vintages are given ratings at a minimum. The absence of ratings is viewed negatively by consumers and can meaningfully impact sales. The firm should also consult with outside marketing experts when outlining new bottle labels. This third-party domain expertise could help safeguard against negative feedback towards future label designs as poor customer reactions in this area could also be detrimental to sales.

## INTRODUCTION

The purpose of this report is to build Poisson and Negative Binomial regression models to predict the number of cases of wine that will be sold given certain properties of the wine. Poisson regression is often used for modeling count data while Negative Binomial regression can be used for over-dispersed count data when a given variance is greater than its mean.

If a wine manufacturer can predict what characteristics are more important to customers, then that manufacturer will be able to adjust their wine offering to maximize sales. We use historical data provided by the course instructor and several software platforms to explore relationships and produce actionable insight. We primarily use functionality within SAS Studio to perform an end-to-end predictive modeling process which utilizes an assortment of continuous and categorical variables to predict wine sales. While the analysis uses several robust, enterprise quality analytics systems, the primary focus continues to be simplicity and interpretability. Given the abundance of continuous variables in this particular data set, we expect to find several chemical properties that have predictive power. We also anticipate at least one of the categorical variables to be useful.

## RESULTS

In the subsequent sections, we generate and evaluate a series of predictive models. We first use the functionality within SAS Studio, Angoss and IBM Watson to perform a brief exploratory data analysis (EDA) to build an understanding of potential predictor variables and their relationship to the response. Following this, we examine the variables for deficiencies such as missing data and outliers as a precursor to preparing the data set for modeling through imputation and elimination. Finally, we construct a series of 5 predictive models. The first 2 models are standard Poisson and Negative Binomial regression models while the next 2 use Zero Inflated Poisson and Negative Binomial distributions. These 4 models will be generated using the PROC GENMOD functionality within SAS Studio. The fifth and final iteration uses standard linear regression which serves as a basis of comparison between the different methodologies. The primary quantitative metrics we will be using for evaluation are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). We then recommend the most logical, effective solution for use by management.

**EXPLORATORY DATA ANALYSIS (EDA).** We use an assortment of tools to perform our initial EDA before cleaning the data and ultimately constructing the predictive models. I begin the analysis by examining the variables in the data set using SAS CONTENT procedure. The data set contains 12,795 observations and 16 variables; 1 variable is a unique identifier (INDEX) and thus excluded from the analysis. The variables are mostly related to the chemical properties of the wine being sold; though several of the variables appear categorical in nature. These variables are AcidIndex, LabelAppeal and STARS. The remaining candidate variables are continuous. The response variable (TARGET) is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases would be used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely a wine is to be sold at a high-end restaurant.

The data set itself has a wide variety of statistics that appear analytically interesting and appropriate to build a model with to predict wine sales. While the data dictionary is usually a helpful resource, it is incomplete in this case. Luckily, the naming convention of the variables is relatively effective. Table 1 below provides an alphabetic list of the possible predictor variables, data types and theoretical effects with TARGET. The theoretical effects are particularly important, where available, as we will reference this logic during the examination of coefficients in the model building process. In the absence of a reliable data dictionary and domain expertise to complete it, we are forced to take the data at face value and infer what effects we can as needed.

**Table 1: Alphabetic List of Variables and Theoretical Effects**

| Variable | Type | Theoretical Effect |
|---|---|---|
| AcidIndex | Categorical | Negative assuming people don't like their wine with a lot of acid |
| Alcohol | Continuous | N/A |
| Chlorides | Continuous | N/A |
| CitricAcid | Continuous | N/A |
| Density | Continuous | N/A |
| FixedAcidity | Continuous | N/A |
| FreeSulfurDioxide | Continuous | N/A |
| LabelAppeal | Categorical | Higher score means more appeal, equating to more sales |
| ResidualSugar | Continuous | N/A |
| STARS | Categorical | Higher rating from experts have a positive impact on sales |
| Sulphates | Continuous | N/A |
| TotalSulfurDioxide | Continuous | N/A |
| VolatileAcidity | Continuous | N/A |
| pH | Continuous | N/A |

**Correlations.** At this juncture, we have posited the effects of the categorical variables but need a better understanding of the relationships between the other possible independent variables and the response. Using Angoss, we begin our analysis of the independent variables using a correlation matrix. Table 2 below provides this output. It is encouraging to see that the 3 categorical variables (AcidIndex, LabelAppeal, STARS) are relatively highly correlated with TARGET and in the predicted directions. The continuous variables are largely uncorrelated with both the dependent variable and among each other. The other meaningful relationship to note is the mild positive correlation (0.33) between STARS and LabelAppeal. This suggests that bottles that are more esthetically appealing are rated higher. This makes sense as label design is likely both a conscious and unconscious factor when assigning ranks to wine.

**(BINGO BONUS #1) Table 2: Correlations with Angoss**

| | TARGET | FixedAcidity | VolatileAcidity | CitricAcid | ResidualSugar | Chlorides | FreeSulfurDioxide | TotalSulfurDioxide | Density | pH | Sulphates | Alcohol | LabelAppeal | AcidIndex | STARS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TARGET | 1 | -0.04901 | -0.08879 | 0.00868 | 0.01649 | -0.03826 | 0.04382 | 0.05148 | -0.03552 | -0.00944 | -0.03885 | 0.06206 | 0.3565 | -0.24605 | 0.55879 |
| FixedAcidity | -0.04901 | 1 | 0.01238 | 0.01424 | -0.01885 | -0.00046 | 0.00497 | -0.0225 | 0.00648 | -0.00898 | 0.03078 | -0.00937 | -0.00337 | 0.17844 | -0.00663 |
| VolatileAcidity | -0.08879 | 0.01238 | 1 | -0.01695 | -0.00648 | 0.00099 | -0.00708 | -0.02108 | 0.01473 | 0.01359 | 0.00013 | 0.00407 | -0.01699 | 0.04464 | -0.03443 |
| CitricAcid | 0.00868 | 0.01424 | -0.01695 | 1 | -0.00694 | -0.00857 | 0.00643 | 0.00632 | -0.01395 | -0.00871 | -0.01299 | 0.01705 | 0.00865 | 0.0657 | 0.00066 |
| ResidualSugar | 0.01649 | -0.01885 | -0.00648 | -0.00694 | 1 | -0.00559 | 0.01749 | 0.02248 | 0.0041 | 0.01212 | -0.00772 | -0.02 | 0.00232 | -0.00941 | 0.01674 |
| Chlorides | -0.03826 | -0.00046 | 0.00099 | -0.00857 | -0.00559 | 1 | -0.02066 | -0.01399 | 0.02266 | -0.01761 | -0.00329 | -0.01969 | 0.01051 | 0.02524 | -0.00493 |
| FreeSulfurDioxide | 0.04382 | 0.00497 | -0.00708 | 0.00643 | 0.01749 | -0.02066 | 1 | 0.01372 | 0.00318 | 0.00605 | 0.01159 | -0.01859 | 0.01029 | -0.04172 | -0.00908 |
| TotalSulfurDioxide | 0.05148 | -0.0225 | -0.02108 | 0.00632 | 0.02248 | -0.01399 | 0.01372 | 1 | 0.01282 | -0.00434 | -0.00713 | -0.01596 | -0.00975 | -0.04931 | 0.01393 |
| Density | -0.03552 | 0.00648 | 0.01473 | -0.01395 | 0.0041 | 0.02266 | 0.00318 | 0.01282 | 1 | 0.00577 | -0.00906 | -0.00721 | -0.00937 | 0.04041 | -0.01828 |
| pH | -0.00944 | -0.00898 | 0.01359 | -0.00871 | 0.01212 | -0.01761 | 0.00605 | -0.00434 | 0.00577 | 1 | 0.00548 | -0.01155 | 0.00414 | -0.05868 | -0.00049 |
| Sulphates | -0.03885 | 0.03078 | 0.00013 | -0.01299 | -0.00772 | -0.00329 | 0.01159 | -0.00713 | -0.00906 | 0.00548 | 1 | 0.00474 | -0.00389 | 0.03445 | -0.01231 |
| Alcohol | 0.06206 | -0.00937 | 0.00407 | 0.01705 | -0.02 | -0.01969 | -0.01859 | -0.01596 | -0.00721 | -0.01155 | 0.00474 | 1 | 0.00103 | -0.03814 | 0.06522 |
| LabelAppeal | 0.3565 | -0.00337 | -0.01699 | 0.00865 | 0.00232 | 0.01051 | 0.01029 | -0.00975 | -0.00937 | 0.00414 | -0.00389 | 0.00103 | 1 | 0.02475 | 0.33479 |
| AcidIndex | -0.24605 | 0.17844 | 0.04464 | 0.0657 | -0.00941 | 0.02524 | -0.04172 | -0.04931 | 0.04041 | -0.05868 | 0.03445 | -0.03814 | 0.02475 | 1 | -0.08626 |
| STARS | 0.55879 | -0.00663 | -0.03443 | 0.00066 | 0.01674 | -0.00493 | -0.00908 | 0.01393 | -0.01828 | -0.00049 | -0.01231 | 0.06522 | 0.33479 | -0.08626 | 1 |

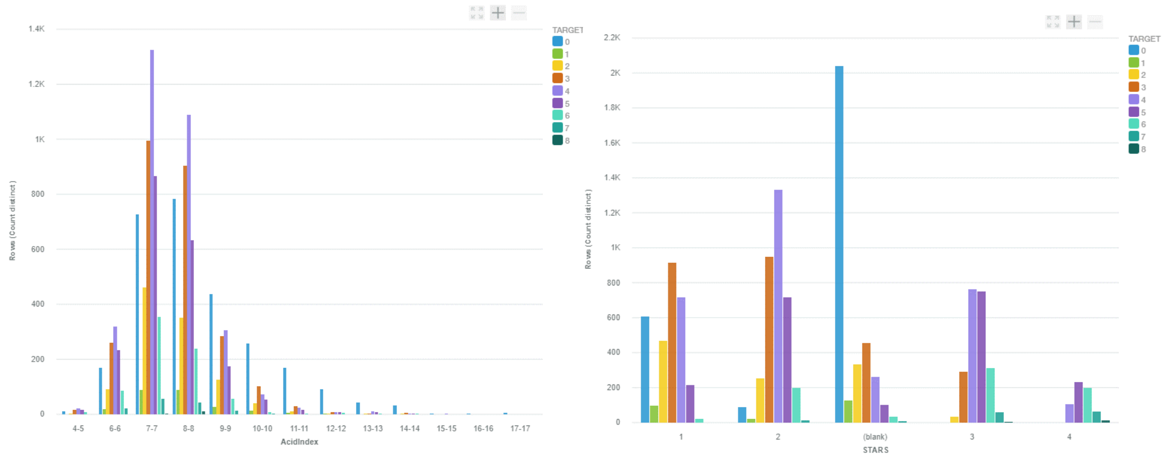
**Identifying Continuous Variables of Interest.** Using the PROC MEANS function with the CLASS set to TARGET_FLAG (a binary variable added to the data set), we decompose the data set further and compare the mean values for all continuous variables by 0 and 1. The objective here is to identify additional predictors despite not having extensive knowledge of the data set. As a rule of thumb, a meaningful difference in average values could signal predictive power. These figures are provided in Table 3 below. While the relative change needs to be considered with regard to the respective variables, several insights can be extracted from the table. There appears to noteworthy mean differences in FreeSulfurDioxide and TotalSulfurDioxide; which makes sense as the two are likely related. Sulfur dioxide is a common preservative used in winemaking and plays two important roles (Godden, Francis, Field, Gishen, Coulter, Valente, Hoj & Robinson, 2002). First, it is an anti-microbial agent, and as such is used to help curtail the growth of undesirable yeasts and bacteria. Secondly, it acts as an antioxidant, safeguarding the wine's fruit integrity and protecting it against browning. We might infer that wines with more (free and/or total) sulfur dioxide look and taste better, resulting in more cases sold.

**Table 3: Mean Analysis of Continuous Variables by TARGET_FLAG**

| Variable | N Miss (0) | N Miss (1) | No Sale (0) | Sale (1) | Delta |
|---|---|---|---|---|---|
| Alcohol | 137.00 | 516.00 | 10.43 | 10.51 | -0.08 |
| Chlorides | 141.00 | 497.00 | 0.08 | 0.05 | 0.03 |
| CitricAcid | 0.00 | 0.00 | 0.30 | 0.31 | -0.01 |
| Density | 0.00 | 0.00 | 1.00 | 0.99 | 0.00 |
| FixedAcidity | 0.00 | 0.00 | 7.73 | 6.90 | 0.84 |
| FreeSulfurDioxide | 139.00 | 508.00 | 17.94 | 34.35 | -16.41 |
| ResidualSugar | 127.00 | 489.00 | 4.00 | 5.81 | -1.81 |
| Sulphates | 279.00 | 931.00 | 0.61 | 0.50 | 0.11 |
| TotalSulfurDioxide | 140.00 | 542.00 | 85.26 | 130.38 | -45.11 |
| VolatileAcidity | 0.00 | 0.00 | 0.45 | 0.29 | 0.16 |
| pH | 101.00 | 294.00 | 3.25 | 3.20 | 0.05 |

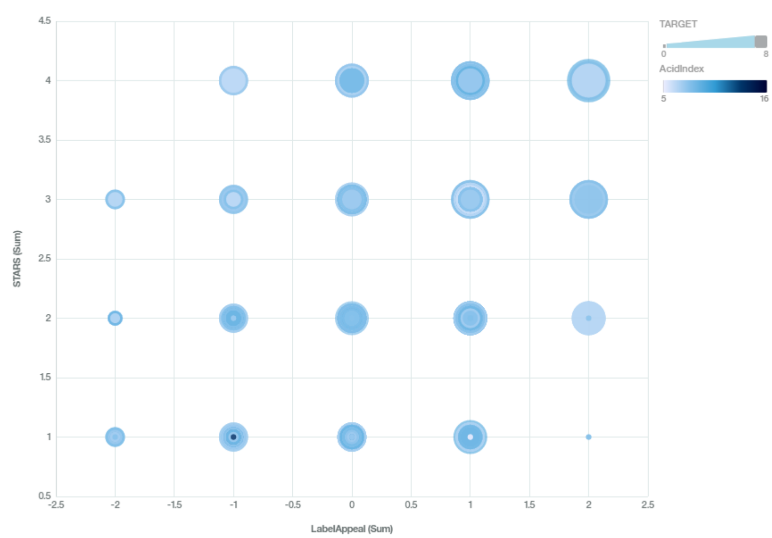
**Identifying Categorical Variables of Interest.** As we have established a relatively firm idea of what continuous variable(s) we will include in the models, we shift our attention to identifying categorical variables of interest (AcidIndex, LabelAppeal and STARS). Using the PROC FREQ function (not shown), we view the categorical variables by TARGET_FLAG and examine which categories sell more cases of wine (i.e. "PROC EYEBALL"). Both AcidIndex and STARS appear analytically interesting. Using IBM Watson Analytics, we generate histograms of each variable by TARGET (Exhibit 2). We can see that people purchase less wine as the amount of acid increases; the most wine appears to be purchased between AcidIndex values of 7 to 8. More interesting is the visual of STARS, which suggests that many people do not purchase any cases if there is no rating. Therefore, missing values of STARS appears highly predictive.

**(BINGO BONUS #2) Exhibit 2: Histograms for AcidIndex and STARS by TARGET**



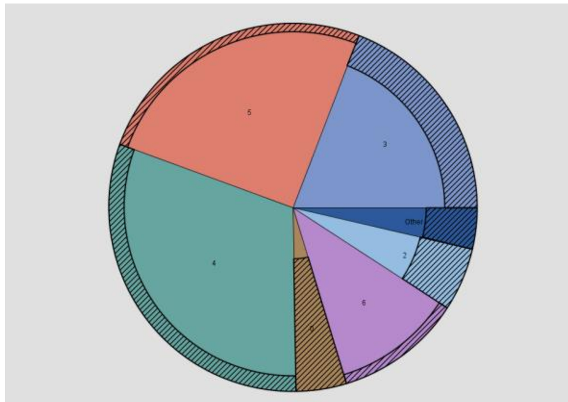

Exploring the relationship between AcidIndex, LabelAppeal and STARS a bit further, we again utilize the discovery tools in IBM Watson Analytics to garner insight into how these variables might simultaneously interact with TARGET (Exhibit 3). For reference, the size of the circles represents the TARGET and the coloring represents AcidIndex. We can see that greater values of STARS (y-axis) and LabelAppeal (x-axis) tend to equate to more cases of wine sold (upper right), which makes sense intuitively. A key threshold appears to be below 2 STARS and 2 LabelAppeal, where no cases of wine are sold. Additionally, no cases of wine are sold at 4 STARS and -2 LabelAppeal, though wines with worse ratings are sold at that same LabelAppeal level, which is counterintuitive. Assuming the price of the wine increases with STARS, sales in the lower left corner could represent the more price sensitive spectrum of the customer base who buy lower rated wine because they are cheaper and are indifferent to the label design. In terms of AcidIndex, there does not seem to be an obvious relationship between it and the other variables. Overall, we can infer that STARS and LabelAppeal are potentially predictive of TARGET.

**(BINGO BONUS #3) Exhibit 3: What is the Relationship Between LabelAppeal and STARS by TARGET?**



Lastly, we utilize functionality in SAS Enterprise Miner to complete the analysis of the categorical variables. Exhibit 4 below provides a pie chart of the number of cases sold (TARGET) with the shaded region representing the proportion of STARS values. We can see that STARS has a meaningful impact when 0 cases are sold. This is in line with earlier observations.

**(BINGO BONUS #4) Exhibit 4: Frequency of STARS by TARGET**



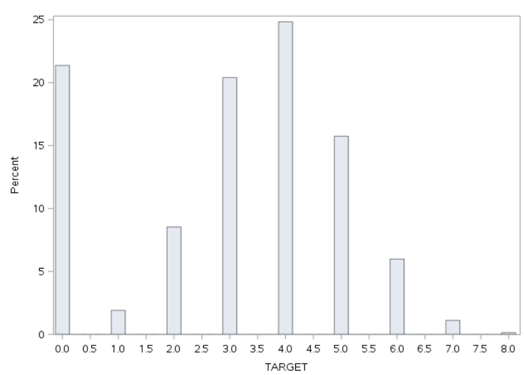

**Analysis of Response Variable.** Given that Poisson regression requires that the variance of a distribution is equal to the mean, this assumption is tested in Table 4 below. We see that TARGET violates the mean and variance assumption for the Poisson distribution, however, it is not in violation for the Negative Bionomial distribution which stipulates the variance to be larger than the mean.

**Table 4: Equality Check**

| Analysis Variable : TARGET | |
|---|---|
| **Mean** | **Variance** |
| 3.0290739 | 3.7108945 |

We further analyze the dependent variable in Exhibit 5 below. TARGET appears to be relatively normally distributed but is likely zero inflated. While in previous projects we have adjusted our approach based on this observation, for purposes of this exercise we will not alter the variable as we are interested in examining the differences in performance across models.

**Exhibit 5: Distribution of TARGET**



**Outliers.** As we begin to transition into the data preparation portion of the modeling process, we take a step back to identify potential outliers in the variables. Table 5 below provides a statistical summary of all variables in the data set. Variables that appear to have possible outliers are AcidIndex, Alcohol, Chlorides, CitricAcid, FixedAcidity, FreeSulfurDioxide, ResidualSugar, Sulphates, TotalSulfurDioxide and VolatileAcidity. Several of the variables have negative numbers but as we don't have domain expertise we cannot comment if this suggests poor data quality. Generally, there could be a number reasons for the presence outliers and poor data quality. First, the original proprietors of the data altered and scaled the data differently than what was initially measured. Additionally, having not collected these data ourselves we are forced to rely on the original practitioners in terms of collection techniques and accuracy. We discuss how to handle the presence of outliers in the subsequent section on data preparation.

**Missing Data.** Similar to identifying outliers, Table 5 assists in finding variables where data is missing. In the wine data set, there are several variables with missing records: Alcohol, Chlorides, FreeSulfurDioxide, ResidualSugar, STARS, Sulphates, TotalSulfurDioxide and pH. STARS in particular is missing almost 25% of the total records.

**Table 5: The MEANS Procedure**

| Variable | N Miss | Mean | Variance | Minimum | 1st Pctl | Median | 99th | Maximum |
|---|---|---|---|---|---|---|---|---|
| AcidIndex | 0 | 7.77 | 1.75 | 4 | 6 | 8 | 13 | 17 |
| Alcohol | 653 | 10.49 | 13.90 | -4.7 | 0.1 | 10.4 | 20.3 | 26.5 |
| Chlorides | 638 | 0.05 | 0.10 | -1.171 | -0.859 | 0.046 | 0.957 | 1.351 |
| CitricAcid | 0 | 0.31 | 0.74 | -3.24 | -2.18 | 0.31 | 2.66 | 3.86 |
| Density | 0 | 0.99 | 0.00 | 0.88809 | 0.9168 | 0.99449 | 1.06981 | 1.09924 |
| FixedAcidity | 0 | 7.08 | 39.91 | -18.1 | -10.9 | 6.9 | 24.4 | 34.4 |
| FreeSulfurDioxide | 647 | 30.85 | 22116.02 | -555 | -388 | 30 | 469 | 623 |
| LabelAppeal | 0 | -0.01 | 0.79 | -2 | -2 | 0 | 2 | 2 |
| ResidualSugar | 616 | 5.42 | 1139.02 | -127.8 | -91 | 3.9 | 99.2 | 141.15 |
| STARS | 3359 | 2.04 | 0.81 | 1 | 1 | 2 | 4 | 4 |
| Sulphates | 1210 | 0.53 | 0.87 | -3.13 | -2.13 | 0.5 | 3.16 | 4.24 |
| TotalSulfurDioxide | 682 | 120.71 | 53783.74 | -823 | -531 | 123 | 767 | 1057 |
| VolatileAcidity | 0 | 0.32 | 0.61 | -2.79 | -1.865 | 0.28 | 2.59 | 3.68 |
| pH | 395 | 3.21 | 0.46 | 0.48 | 1.32 | 3.2 | 5.125 | 6.13 |

The EDA uncovered several interesting elements in the data set. The categorical variables generally seem more predictive than the continuous variables. Also, the fact that there are a large number of missing STAR variables seems highly predictive. Overall, we successfully identified several variables manually which will augment our automated variable selection in subsequent sections. Below is a summary of the variables we posit may have predictive power and will keep under consideration throughout the model building process:

- ***Continuous Variables:*** FreeSulfurDioxide and TotalSulfurDioxide

- ***Categorical Variables:*** AcidIndex, LabelAppeal and STARS

**DATA PREPARATION.** In this section, we prepare the data by changing the original values as well as create several new variables. Before moving on to the model portion of the discussion, we present a summary of the new data set.

**Outlier Resolution**. In attempt to create a better fit for our models, we first modify the data set to eliminate possible outliers as they can significantly influence results. To accomplish this, we impute outlier values, replacing them with the 1$^{st}$ and 99$^{th}$ percentile breakpoints for each respective variable. The 99$^{th}$ percentile breakpoint is applied to AcidIndex and FixedAcidity as they have extreme high values. The majority of the variables have both extreme low and high values so we impute at both breakpoints in the case of the following: Alcohol, Chlorides, CitricAcid, FreeSulfurDioxide, ResidualSugar, Sulphates, TotalSulfurDioxide and VolatileAcidity.

**Missing Data Resolution.** Similar to the outlier remedy, we impute missing values in the data set by replacing them with the median values of the distribution. While there are more sophisticated replacement techniques, median imputation is used because it is a number that is already present in the data set and is less susceptible to outlier errors as compared to mean imputation. In addition to median imputation, a missing flag for each variable is generated to determine whether there is a difference in outcomes associated with missing versus complete data. We apply this methodology to Alcohol, Chlorides, FreeSulfurDioxide, ResidualSugar, STARS, Sulphates, TotalSulfurDioxide and pH.

The resultant structure of the new data set (not shown) has no missing values, 11 imputed variables, 7 missing data flags and 2 original variables. Outlier deletion and imputation are intended to improve the robustness of models and can have powerful effects on fit. While the effects are sometimes not in the desired direction, these analytical activities can be beneficial if they improve the relationship between two or more variables.

**BUILD MODELS.** In this section, we generate 5 regression models and discuss the findings. The first 2 models use PROC GENMOD with Poisson and Negative Binomial distributions, respectively. The next 2 models use PROC GENMOD with a Zero Inflated Poisson and Negative Binomial distributions, respectively. The final model uses PROC REG to produce a standard regression model as a basis of comparison to other techniques.

**Variable Selection.** Before presenting the predictive models, we briefly discuss the variable selection technique used. Table 6 below shows the variables selected by 4 different analytics platforms by order of importance for predicting TARGET. For reference, PROC HPSPLIT was also used to select variables for the 'zero models' in Models 3 and 4 (Not shown / **BINGO BONUS #6**).

For purposes of predicting TARGET, we use the variables from PROC HPSPLIT (i.e. a decision tree) in SAS Studio for Models 1 through 4 while Model 5 uses stepwise selection. Enterprise Miner provides too few variables while Azure provides too many. We are also not satisfied with the rankings from the Angoss decision tree as the highly ranked variables a relatively different than the other selection techniques. Lastly, it is encouraging that 3 of the 4 manually identified variables were selected by the SAS Studio decision tree.

**(BINGO BONUS # 5) Table 6: TARGET Variable Selection**

| SAS Studio Decision Tree | Enterprise Miner | Microsoft Azure ML | Angoss Decision Tree |
|---|---|---|---|
| M STARS | IMP STARS | M STARS | LabelAppeal |
| LabelAppeal | LabelAppeal | IMP AcidIndex | IMP Alcohol |
| IMP STARS | IMP AcidIndex | IMP VolatileAcidity | IMP Chlorides |
| IMP Alcohol | IMP VolatileAcidity | IMP Chlorides | IMP ResidualSugar |
| IMP AcidIndex | | IMP Sulphates | IMP STARS |
| IMP VolatileAcidity | | IMP ph | M STARS |
| IMP TotalSulfurDioxide | | Density | IMP TotalSulfurDioxide |
| IMP Chlorides | | M IMP ph | IMP ph |
| IMP ResidualSugar | | M Sulphates | IMP AcidIndex |
| IMP ph | | IMP ResidualSugar | IMP CitricAcid |
| Density | | M Alcohol | IMP VolatileAcidity |
| IMP FixedAcidity | | M FreeSulfurDioxide | |
| IMP CitricAcid | | M TotalSulfurDioxide | |
| | | M ResidualSugar | |
| | | IMP CitricAcid | |
| | | IMP Alcohol | |
| | | IMP FreeSulfurDioxide | |
| | | IMP TotalSulfurDioxide | |
| | | IMP STARS | |
| | | LabelAppeal | |

**Model 1: Poisson.** Using the SAS GENMOD regression procedure and a Poisson distribution, we generate the following results (Table 7 and Table 8). Note we do not include IMP_AcidIndex in the CLASS statement for any of the models as it would make presenting the results unwieldy:

**Table 7: Poisson Model Analysis Of Maximum Likelihood Parameter Estimates**

| Parameter | Set | Estimate | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Intercept | | 1.6288 | 67.48 | <.0001 |
| M_STARS | 0 | 1.0867 | 3550.63 | <.0001 |
| M_STARS | 1 | 0 | . | . |
| LabelAppeal | -2 | -0.6964 | 269.28 | <.0001 |
| LabelAppeal | -1 | -0.4603 | 339.56 | <.0001 |
| LabelAppeal | 0 | -0.27 | 139.46 | <.0001 |
| LabelAppeal | 1 | -0.1378 | 35.36 | <.0001 |
| LabelAppeal | 2 | 0 | . | . |
| IMP_STARS | 1 | -0.5579 | 663.31 | <.0001 |
| IMP_STARS | 2 | -0.2388 | 144.07 | <.0001 |
| IMP_STARS | 3 | -0.1193 | 34.86 | <.0001 |
| IMP_STARS | 4 | 0 | . | . |
| IMP_Alcohol | | 0.0039 | 7.03 | 0.008 |
| IMP_AcidIndex | | -0.0814 | 308.4 | <.0001 |
| IMP_VolatileAcidity | | -0.0315 | 21.73 | <.0001 |
| IMP_TotalSulfurDioxide | | 0.0001 | 12.11 | 0.0005 |
| IMP_Chlorides | | -0.0388 | 5.22 | 0.0223 |
| IMP_ResidualSugar | | 0.0001 | 0.27 | 0.6012 |
| IMP_ph | | -0.0125 | 2.66 | 0.1027 |
| Density | | -0.2557 | 1.78 | 0.1825 |
| IMP_FixedAcidity | | -0.0001 | 0.01 | 0.9097 |
| IMP_CitricAcid | | 0.0059 | 0.92 | 0.3374 |

**Table 8: Poisson Model Criteria For Assessing Goodness Of Fit**

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | 1.30E+04 | 13649.285 | 1.0684 |
| Scaled Deviance | 1.30E+04 | 13649.285 | 1.0684 |
| Pearson Chi-Square | 1.30E+04 | 11277.569 | 0.8827 |
| Scaled Pearson X2 | 1.30E+04 | 11277.569 | 0.8827 |
| Log Likelihood | | 8801.5179 | |
| Full Log Likelihood | | -22795.653 | |
| AIC (smaller is better) | | 45629.3067 | |
| AICC (smaller is better) | | 45629.3662 | |
| BIC (smaller is better) | | 45770.9861 | |

To begin the interpretation of Model 1, if a type of wine is not given a STAR ranking, the model suggests a significant decrease (196%) in the expected number of cases to be sold.

Given that LabelAppeal has a base level of 2, we interpret the coefficients as being rated a:

- -2: 50% decrease in the expected number of cases sold
- -1: 37% decrease in the expected number of cases sold
- 0: 24% decrease in the expected number of cases sold
- +1: 13% decrease in the expected number of cases sold

Given that IMP_STARS has a base level of 4, we interpret the coefficients as being rated a:

- 1: 43% decrease in the expected number of cases sold
- 2: 21% decrease in the expected number of cases sold
- 3: 11% decrease in the expected number of cases sold

In terms of the continuous variables, we only interpret one for the sake of brevity. The effect of a one-unit increase in Density equates to a 23% decrease in the expected number of cases purchased.

The initial assessment of Model 1 is that the categorical variables are in the predicted direction and significant and the one continuous variable (IMP_TotalSulfurDioxide) we identified was significant and the correct sign. The majority of continuous variables are generally not statistically significant so we would likely remove them in future iterations however that is beyond the scope of this exercise.

**Model 2: Negative Binomial.** For the second model, we use the SAS GENMOD regression procedure and a Negative Binomial distribution. These results are presented in Table 9 and Table 10 below:

**Table 9: NB Model Analysis Of Maximum Likelihood Parameter Estimates**

| Parameter | Set | Estimate | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Intercept | | 1.6288 | 67.48 | <.0001 |
| M_STARS | 0 | 1.0867 | 3550.63 | <.0001 |
| M_STARS | 1 | 0 | . | . |
| LabelAppeal | -2 | -0.6964 | 269.28 | <.0001 |
| LabelAppeal | -1 | -0.4603 | 339.55 | <.0001 |
| LabelAppeal | 0 | -0.27 | 139.46 | <.0001 |
| LabelAppeal | 1 | -0.1378 | 35.36 | <.0001 |
| LabelAppeal | 2 | 0 | . | . |
| IMP_STARS | 1 | -0.5579 | 663.31 | <.0001 |
| IMP_STARS | 2 | -0.2388 | 144.07 | <.0001 |
| IMP_STARS | 3 | -0.1193 | 34.86 | <.0001 |
| IMP_STARS | 4 | 0 | . | . |
| IMP_Alcohol | | 0.0039 | 7.03 | 0.008 |
| IMP_AcidIndex | | -0.0814 | 308.39 | <.0001 |
| IMP_VolatileAcidity | | -0.0315 | 21.73 | <.0001 |
| IMP_TotalSulfurDioxi | | 0.0001 | 12.11 | 0.0005 |
| IMP_Chlorides | | -0.0388 | 5.22 | 0.0223 |
| IMP_ResidualSugar | | 0.0001 | 0.27 | 0.6012 |
| IMP_ph | | -0.0125 | 2.66 | 0.1027 |
| Density | | -0.2557 | 1.78 | 0.1825 |
| IMP_FixedAcidity | | -0.0001 | 0.01 | 0.9097 |
| IMP_CitricAcid | | 0.0059 | 0.92 | 0.3374 |

**Table 10: NB Model Criteria For Assessing Goodness Of Fit**

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | 1.30E+04 | 13649.285 | 1.0684 |
| Scaled Deviance | 1.30E+04 | 13649.285 | 1.0684 |
| Pearson Chi-Square | 1.30E+04 | 11277.569 | 0.8827 |
| Scaled Pearson X2 | 1.30E+04 | 11277.569 | 0.8827 |
| Log Likelihood | | 8801.5179 | |
| Full Log Likelihood | | -22795.6533 | |
| AIC (smaller is better) | | 45629.3067 | |
| AICC (smaller is better) | | 45629.3662 | |
| BIC (smaller is better) | | 45770.9861 | |

Model 2 is very similar to Model 1 due to the similarities in mean and variance, therefore we will not repeat the interpretation of the coefficients.

**Model 3: Zero Inflated Poisson (ZIP).** Using the SAS GENMOD regression procedure, a Poisson distribution and a zero model, we generate the following results (Tables 11 through 13):

**Table 11: ZIP Model Analysis Of Maximum Likelihood Parameter Estimates**

| Parameter | Set | Estimate | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Intercept | | 2.0459 | 101.18 | <.0001 |
| M_STARS | 0 | 0.1828 | 85.52 | <.0001 |
| M_STARS | 1 | 0 | . | . |
| LabelAppeal | -2 | -1.0773 | 557.82 | <.0001 |
| LabelAppeal | -1 | -0.6363 | 616.86 | <.0001 |
| LabelAppeal | 0 | -0.348 | 225.72 | <.0001 |
| LabelAppeal | 1 | -0.1579 | 45.65 | <.0001 |
| LabelAppeal | 2 | 0 | . | . |
| IMP_STARS | 1 | -0.3171 | 205.38 | <.0001 |
| IMP_STARS | 2 | -0.1956 | 95.7 | <.0001 |
| IMP_STARS | 3 | -0.0982 | 23.62 | <.0001 |
| IMP_STARS | 4 | 0 | . | . |
| IMP_Alcohol | | 0.007 | 21.96 | <.0001 |
| IMP_AcidIndex | | -0.0199 | 16.15 | <.0001 |
| IMP_VolatileAcidity | | -0.0123 | 3.14 | 0.0763 |
| IMP_TotalSulfurDioxide | | 0 | 0.43 | 0.5142 |
| IMP_Chlorides | | -0.0259 | 2.22 | 0.1359 |
| IMP_ResidualSugar | | 0 | 0.06 | 0.8119 |
| IMP_ph | | 0.0023 | 0.09 | 0.7641 |
| Density | | -0.2637 | 1.8 | 0.1801 |
| IMP_FixedAcidity | | 0.0003 | 0.12 | 0.7251 |
| IMP_CitricAcid | | 0.0015 | 0.06 | 0.8067 |

**Table 12: ZIP Model Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates**

| Parameter | Set | Estimate | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Intercept | | -21.2536 | 5259.87 | <.0001 |
| M_STARS | 0 | -5.7345 | 293.44 | <.0001 |
| M_STARS | 1 | 0 | . | . |
| IMP_AcidIndex | | 0.4389 | 287.5 | <.0001 |
| IMP_STARS | 1 | 23.0651 | 4708.35 | <.0001 |
| IMP_STARS | 2 | 19.4109 | . | . |
| IMP_STARS | 3 | 0.2251 | 0 | 1 |
| IMP_STARS | 4 | 0 | . | . |
| LabelAppeal | -2 | -3.3582 | 76.33 | <.0001 |
| LabelAppeal | -1 | -1.8823 | 74.99 | <.0001 |
| LabelAppeal | 0 | -1.1353 | 28.99 | <.0001 |
| LabelAppeal | 1 | -0.4322 | 3.95 | 0.0468 |
| LabelAppeal | 2 | 0 | . | . |
| IMP_VolatileAcidity | | 0.196 | 19.09 | <.0001 |
| IMP_TotalSulfurDioxi | | -0.001 | 42.05 | <.000 |

**Table 13: ZIP Model Criteria For Assessing Goodness Of Fit**

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | | 40726.2302 | |
| Scaled Deviance | | 40726.2302 | |
| Pearson Chi-Square | 1.30E+04 | 5680.556 | 0.445 |
| Scaled Pearson X2 | 1.30E+04 | 5680.556 | 0.445 |
| Log Likelihood | | 11234.0562 | |
| Full Log Likelihood | | -20363.1151 | |
| AIC (smaller is better) | | 40788.2302 | |
| AICC (smaller is better) | | 40788.3856 | |
| BIC (smaller is better) | | 41019.3913 | |

To begin the interpretation of Model 3, if a type of wine is not given a STAR ranking, the model suggests a 20% decrease in the expected number of cases to be sold.

Given that LabelAppeal has a base level of 2, we interpret the coefficients as being rated a:

- -2: 66% decrease in the expected number of cases sold
- -1: 47% decrease in the expected number of cases sold
- 0: 29% decrease in the expected number of cases sold
- +1: 15% decrease in the expected number of cases sold

Given that IMP_STARS has a base level of 4, we interpret the coefficients as being rated a:

- 1: 27% decrease in the expected number of cases sold
- 2: 18% decrease in the expected number of cases sold
- 3: 9% decrease in the expected number of cases sold

In terms of the continuous variables, we only interpret one. The effect of a one-unit increase in Density equates to a 23% decrease in the expected number of cases purchased, the same as Models 1 and 2.

The zero inflation portion of the model uses logistic regression to predict the likelihood that the number of cases is 0 (i.e. not sold). If a type of wine does not have a rating there is a 99% increase in the odds that the wine will not be purchased. The interpretation for the remaining variables is similar and thus only interpret one for the sake of brevity.

The initial assessment of Model 3 is that the categorical variables are in the predicted direction and significant and the one continuous variable (IMP_TotalSulfurDioxide) we identified was not significant and had no impact. The majority of continuous variables were again not statistically significant, corroborating the earlier notion of removing them from future iterations.

**Model 4: Zero Inflated Negative Binomial (ZINB).** Using the SAS GENMOD regression procedure, a Negative Binomial distribution and a zero model, we generate the following results (Tables 14 through16):

**Table 14: ZINB Model Analysis Of Maximum Likelihood Parameter Estimates**

| Parameter | Set | Estimate | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Intercept | | 2.047 | 100.52 | <.0001 |
| M_STARS | 0 | 0.1824 | 85.39 | <.0001 |
| M_STARS | 1 | 0 | . | . |
| LabelAppeal | -2 | -1.073 | 557.95 | <.0001 |
| LabelAppeal | -1 | -0.6373 | 615.36 | <.0001 |
| LabelAppeal | 0 | -0.3491 | 225.84 | <.0001 |
| LabelAppeal | 1 | -0.1583 | 45.61 | <.0001 |
| LabelAppeal | 2 | 0 | . | . |
| IMP_STARS | 1 | -0.3156 | 201.84 | <.0001 |
| IMP_STARS | 2 | -0.1913 | 91.02 | <.0001 |
| IMP_STARS | 3 | -0.0981 | 23.35 | <.0001 |
| IMP_STARS | 4 | 0 | . | . |
| IMP_Alcohol | | 0.007 | 21.83 | <.0001 |
| IMP_AcidIndex | | -0.0187 | 14.22 | 0.0002 |
| IMP_VolatileAcidity | | -0.0119 | 2.92 | 0.0874 |
| IMP_TotalSulfurDioxi | | 0 | 0.53 | 0.4646 |
| IMP_Chlorides | | -0.0248 | 2.04 | 0.1534 |
| IMP_ResidualSugar | | 0 | 0.08 | 0.7818 |
| IMP_ph | | 0.0028 | 0.13 | 0.7206 |
| Density | | -0.2744 | 1.93 | 0.1649 |
| IMP_FixedAcidity | | 0.0003 | 0.15 | 0.6956 |
| IMP_CitricAcid | | 0.0011 | 0.03 | 0.8574 |

**Table 15: ZINB Model Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates**

| Parameter | Set | Estimate | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Intercept | | -2.8464 | 29 | <.0001 |
| M_STARS | 0 | -4.647 | 834.72 | <.0001 |
| M_STARS | 1 | 0 | . | . |
| IMP_AcidIndex | | 0.4179 | 296.11 | <.0001 |
| IMP_STARS | 1 | 3.6253 | 66.03 | <.0001 |
| IMP_STARS | 2 | 1.023 | 4.81 | 0.0283 |
| IMP_STARS | 3 | 0.2061 | 0.17 | 0.6784 |
| IMP_STARS | 4 | 0 | . | . |
| LabelAppeal | -2 | -3.0125 | 82.42 | <.0001 |
| LabelAppeal | -1 | -1.7098 | 77.36 | <.0001 |
| LabelAppeal | 0 | -0.9891 | 27.98 | <.0001 |
| LabelAppeal | 1 | -0.3445 | 3.22 | 0.0727 |
| LabelAppeal | 2 | 0 | . | . |
| IMP_VolatileAcidity | | 0.1892 | 19.37 | <.0001 |
| IMP_TotalSulfurDioxi | | -0.001 | 41.36 | <.0001 |
| Intercept | | -2.8464 | 29 | <.0001 |

**Table 16: ZINB Model Criteria For Assessing Goodness Of Fit**

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | | 40853.0116 | |
| Scaled Deviance | | 40853.0116 | |
| Pearson Chi-Square | 1.30E+04 | 5523.758 | 0.4328 |
| Scaled Pearson X2 | 1.30E+04 | 5523.758 | 0.4328 |
| Log Likelihood | | -20426.5058 | |
| Full Log Likelihood | | -20426.5058 | |
| AIC (smaller is better) | | 40917.0116 | |
| AICC (smaller is better) | | 40917.1771 | |
| BIC (smaller is better) | | 41155.6295 | |

Model 4 is very similar to Model 3 despite a notable difference in Intercept values, therefore, we will not repeat the interpretation of the coefficients.

**Model 5: Linear Regression.** For the fifth and final model, we use the PROC REG procedure for linear regression and stepwise selection. These results are presented in Table 17 and Table 18 below:

**Table 17: Parameter Estimates for Linear Regression**

| Variable | Parameter Estimate | Standard Error | F Value | Pr > F | VIF |
|---|---|---|---|---|---|
| Intercept | 4.4531 | 0.44503 | 100.13 | <.0001 | 0 |
| Density | -0.7939 | 0.43715 | 3.3 | 0.0694 | 1.00322 |
| LabelAppeal | 0.46657 | 0.01367 | 1165.27 | <.0001 | 1.10577 |
| IMP_Alcohol | 0.0125 | 0.00331 | 14.25 | 0.0002 | 1.00637 |
| IMP_Chlorides | -0.11967 | 0.03856 | 9.63 | 0.0019 | 1.00267 |
| IMP_STARS | 0.77826 | 0.01568 | 2464.76 | <.0001 | 1.10103 |
| M_STARS | -2.24462 | 0.02696 | 6934.09 | <.0001 | 1.04874 |
| IMP_Sulphates | -0.03137 | 0.01352 | 5.39 | 0.0203 | 1.00215 |
| IMP_TotalSulfurDioxide | 0.00023137 | 0.00005315 | 18.95 | <.0001 | 1.00418 |
| IMP_ph | -0.03191 | 0.01735 | 3.38 | 0.0659 | 1.00488 |
| IMP_AcidIndex | -0.20984 | 0.00926 | 513 | <.0001 | 1.05421 |
| IMP_CitricAcid | 0.0203 | 0.01397 | 2.11 | 0.1461 | 1.00615 |
| IMP_VolatileAcidity | -0.09955 | 0.01534 | 42.1 | <.000 | 1.00642 |

**Table 18: Linear Regression ANOVA Table**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 12 | 25540 | 2128.31 | 1240.1 | <.0001 |
| Error | 12782 | 21938 | 1.71628 | | |
| Corrected Total | 12794 | 47477 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 1.31007 | R-Square | 0.5379 |
| AIC | 6926.37 | Adj R-Sq | 0.5375 |
| BIC | 6924.35 | | |

In terms of Model 5, all of the variables we identified during the EDA are in the predicted direction and statistically significant. The interpretation of Model 5 is simple given the absence of transformations. Within the context of this model, if all of the independent variables were equal to 0, then we expect 4 cases to be sold, which is reasonable given the mean of 3 for TARGET. Additionally, for every increase of 1 unit across all of the independent variables, we expect the average number of cases sold to decrease to 2. The F Value is relatively large and statistically significant. R-Squared and Adjusted R-Squared are 55% and 51%, respectively, which is reasonable. Multicollinearity is not an issue. Overall, the diagnostics suggest Model 5 is sound.

**MODEL SELECTION.** We have presented 5 different predictive models and now must chose the "Best Model." For purposes of this exercise we used a combination of quantitative and qualitive measures to assess the best model. Table 19 below provides metrics for each of the 5 models. The main quantitative criteria we will be evaluating on are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Our guiding rule is to minimize the AIC and BIC. It is important to note that AIC and BIC metrics cannot be compared between count models (i.e. Poisson, NB, etc.) and standard linear regression models.

**Table 19: Model Comparison Statistics**

| Criteria | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| *Method* | *Poisson* | *NB* | *ZIP* | *ZINB* | *OLS* |
| **AIC** | 45629.3067 | 45629.3067 | 40788.2302 | 40917.0116 | 6926.37 |
| **BIC** | 45770.9861 | 45770.9861 | 41019.3913 | 41155.6295 | 6924.35 |

The different regression techniques produced highly similar results between Model 1 and Model 2 as well as Model 3 and Model 4. In terms of the Poisson and Negative Binomial distributions, all of the models are relatively sound, however the zero inflated models appear superior. The outsized impact of M_STAR in Model 1 and Model 2 brings up concern over the quality of the models. Between Model 3 and Model 4, the ZIP model metrics are slightly more appealing, however, recall that the distribution of TARGET violates the assumptions of a Poisson distribution. *For this reason, the recommendation is to use Model 4 or the Zero Inflated Negative Binomial model. This is the iteration that has been submitted to Kaggle for deployment*. Lastly, from a business standpoint, we find the standard linear regression model (Model 5) the simplest to explain given that interpretation of the coefficients is in cases of wine instead of percent change. While linear regression might not be completely appropriate given the structure of this data set, if the results makes sense to decisions makers - use it.

## CONCLUSION

Building sound regression models is a cornerstone of predictive analytics and to creating a culture of proactive data-driven decision making. For this project, a data set containing quantitative and qualitative properties of wine was used to generate Poisson regression models to predict wines sales. Predicting preference in this case is inherently difficult given the subjectivity of individual taste but the potential value for the winemaking industry is evident. The results suggest, at least within the context of this data set, that casual ambiguity and brand management may be more important to selling wine than its chemical properties. This is slightly contrary to our original expectations as we thought a wine's chemical compounds would have had more predictive power. We also did not anticipate missing variables to be so highly predictive. Overall, this exercise reinforces the notion that not all data is equally valuable and there are insights to be drawn from both categorical and continuous variables.

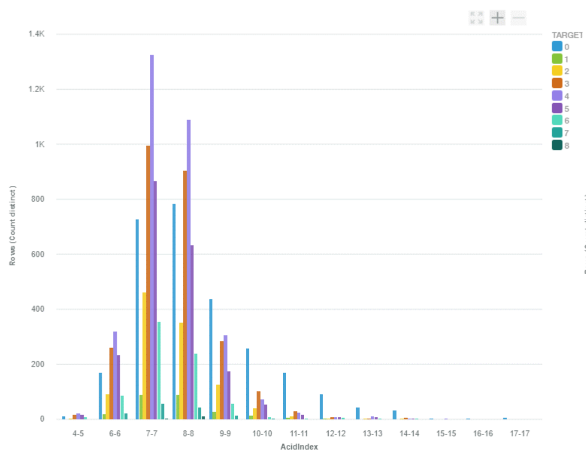
## BINGO BONUS SECTION

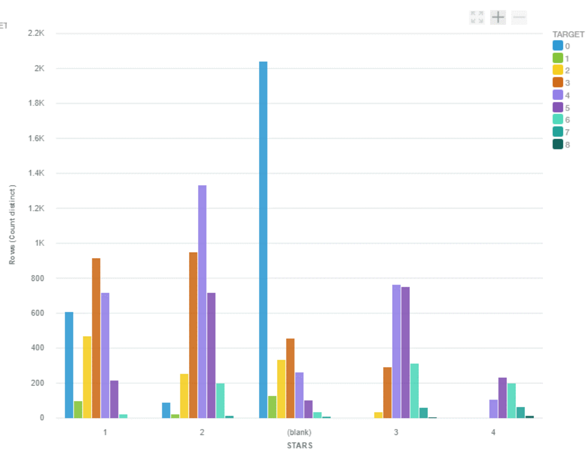
### BINGO BONUS #1 <5 points>: Table 2: Correlations with Angoss

| | TARGET | FixedAcidity | VolatileAcidity | CitricAcid | ResidualSugar | Chlorides | FreeSulfurDioxide | TotalSulfurDioxide | Density | pH | Sulphates | Alcohol | LabelAppeal | AcidIndex | STARS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TARGET | 1 | -0.04901 | -0.08879 | 0.00868 | 0.01649 | -0.03826 | 0.04382 | 0.05148 | -0.03552 | -0.00944 | -0.03885 | 0.06206 | 0.3565 | -0.24605 | 0.55879 |
| FixedAcidity | -0.04901 | 1 | 0.01238 | 0.01424 | -0.01885 | -0.00046 | 0.00497 | -0.0225 | 0.00648 | -0.00898 | 0.03078 | -0.00937 | -0.00337 | 0.17844 | -0.00663 |
| VolatileAcidity | -0.08879 | 0.01238 | 1 | -0.01695 | -0.00648 | 0.00099 | -0.00708 | -0.02108 | 0.01473 | 0.01359 | 0.00013 | 0.00407 | -0.01699 | 0.04464 | -0.03443 |
| CitricAcid | 0.00868 | 0.01424 | -0.01695 | 1 | -0.00694 | -0.00857 | 0.00643 | 0.00632 | -0.01395 | -0.00871 | -0.01299 | 0.01705 | 0.00865 | 0.0657 | 0.00066 |
| ResidualSugar | 0.01649 | -0.01885 | -0.00648 | -0.00694 | 1 | -0.00559 | 0.01749 | 0.02248 | 0.0041 | 0.01212 | -0.00772 | -0.02 | 0.00232 | -0.00941 | 0.01674 |
| Chlorides | -0.03826 | -0.00046 | 0.00099 | -0.00857 | -0.00559 | 1 | -0.02066 | -0.01399 | 0.02266 | -0.01761 | -0.00329 | -0.01969 | 0.01051 | 0.02524 | -0.00493 |
| FreeSulfurDioxide | 0.04382 | 0.00497 | -0.00708 | 0.00643 | 0.01749 | -0.02066 | 1 | 0.01372 | 0.00318 | 0.00605 | 0.01159 | -0.01859 | 0.01029 | -0.04172 | -0.00908 |
| TotalSulfurDioxide | 0.05148 | -0.0225 | -0.02108 | 0.00632 | 0.02248 | -0.01399 | 0.01372 | 1 | 0.01282 | -0.00434 | -0.00713 | -0.01596 | -0.00975 | -0.04931 | 0.01393 |
| Density | -0.03552 | 0.00648 | 0.01473 | -0.01395 | 0.0041 | 0.02266 | 0.00318 | 0.01282 | 1 | 0.00577 | -0.00906 | -0.00721 | -0.00937 | 0.04041 | -0.01828 |
| pH | -0.00944 | -0.00898 | 0.01359 | -0.00871 | 0.01212 | -0.01761 | 0.00605 | -0.00434 | 0.00577 | 1 | 0.00548 | -0.01155 | 0.00414 | -0.05868 | -0.00049 |
| Sulphates | -0.03885 | 0.03078 | 0.00013 | -0.01299 | -0.00772 | -0.00329 | 0.01159 | -0.00713 | -0.00906 | 0.00548 | 1 | 0.00474 | -0.00389 | 0.03445 | -0.01231 |
| Alcohol | 0.06206 | -0.00937 | 0.00407 | 0.01705 | -0.02 | -0.01969 | -0.01859 | -0.01596 | -0.00721 | -0.01155 | 0.00474 | 1 | 0.00103 | -0.03814 | 0.06522 |
| LabelAppeal | 0.3565 | -0.00337 | -0.01699 | 0.00865 | 0.00232 | 0.01051 | 0.01029 | -0.00975 | -0.00937 | 0.00414 | -0.00389 | 0.00103 | 1 | 0.02475 | 0.33479 |
| AcidIndex | -0.24605 | 0.17844 | 0.04464 | 0.0657 | -0.00941 | 0.02524 | -0.04172 | -0.04931 | 0.04041 | -0.05868 | 0.03445 | -0.03814 | 0.02475 | 1 | -0.08626 |
| STARS | 0.55879 | -0.00663 | -0.03443 | 0.00066 | 0.01674 | -0.00493 | -0.00908 | 0.01393 | -0.01828 | -0.00049 | -0.01231 | 0.06522 | 0.33479 | -0.08626 | 1 |

### BINGO BONUS #2 <5 points>: Histograms for AcidIndex and STARS by TARGET

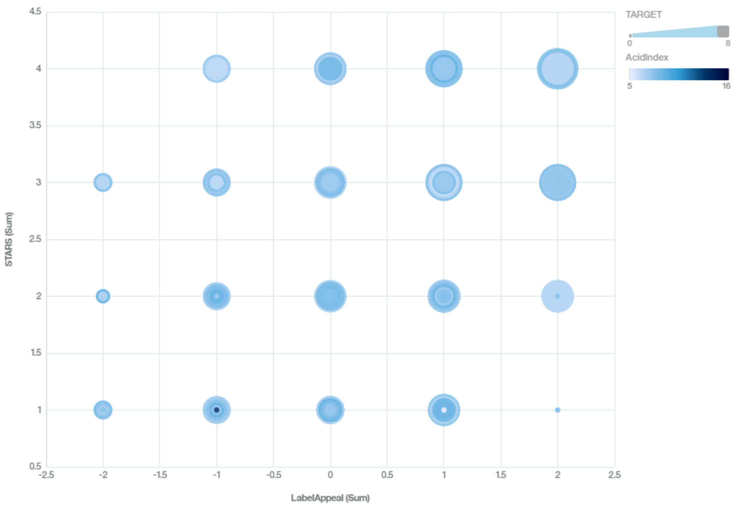
Created with IBM Watson. Useful tool, great visualizations and having the software ask me what I'd like to know was just crazy.



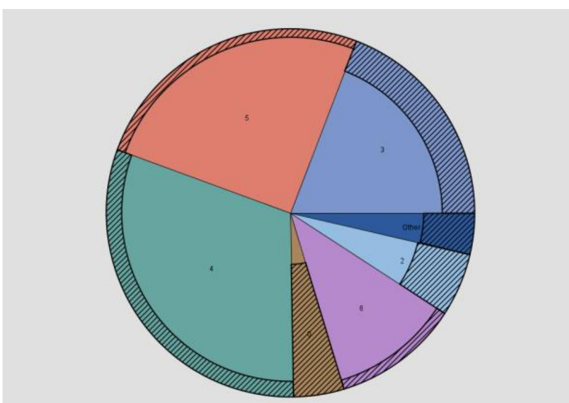
**BINGO BONUS #3 < 5 points>: What is the Relationship Between LabelAppeal and STARS by TARGET?**

Another IBM Watson visual for my EDA.



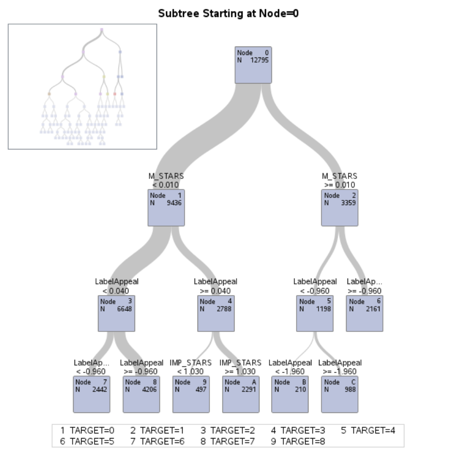

**BINGO BONUS #4 <5 points> Frequency of STARS by TARGET**

Another great sample from SAS Enterprise Miner.



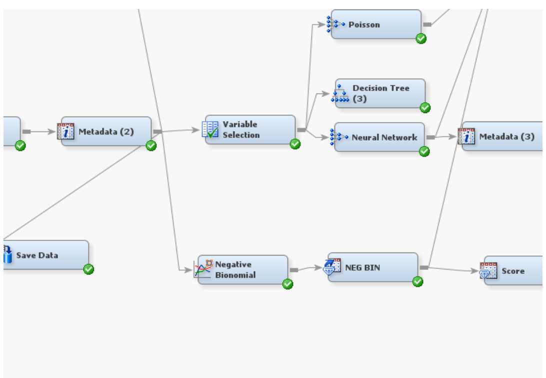
**BINGO BONUS #5 <40 points>:** Used 4 different analytics platforms for TARGET variable selection.

## SAS Studio Decision Tree (PROC HPSPLIT)



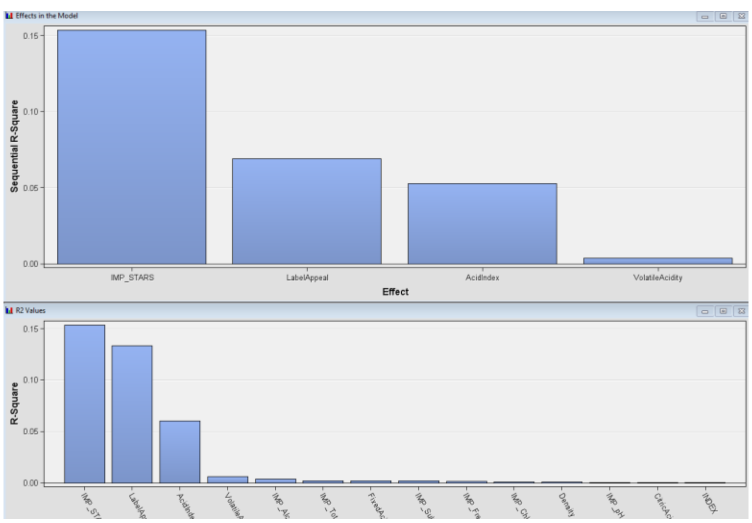

| Variable Importance | | | |
|---|---|---|---|
| | Training | | |
| Variable | Relative | Importance | Count |
| M_STARS | 1.0000 | 30.7564 | 1 |
| LabelAppeal | 0.9660 | 29.7096 | 7 |
| IMP_STARS | 0.5939 | 18.2674 | 8 |
| IMP_Alcohol | 0.3008 | 9.2508 | 11 |
| IMP_AcidIndex | 0.2789 | 8.5790 | 6 |
| IMP_VolatileAcidity | 0.1259 | 3.8714 | 2 |
| IMP_TotalSulfurDioxide | 0.1230 | 3.7818 | 3 |
| IMP_Chlorides | 0.1061 | 3.2623 | 2 |
| IMP_ResidualSugar | 0.0898 | 2.7625 | 1 |
| IMP_ph | 0.0849 | 2.6122 | 1 |
| Density | 0.0532 | 1.6354 | 1 |
| IMP_FixedAcidity | 0.0436 | 1.3425 | 1 |
| IMP_CitricAcid | 0.0368 | 1.1324 | 1 |

## SAS Enterprise Miner Variable Selection Node



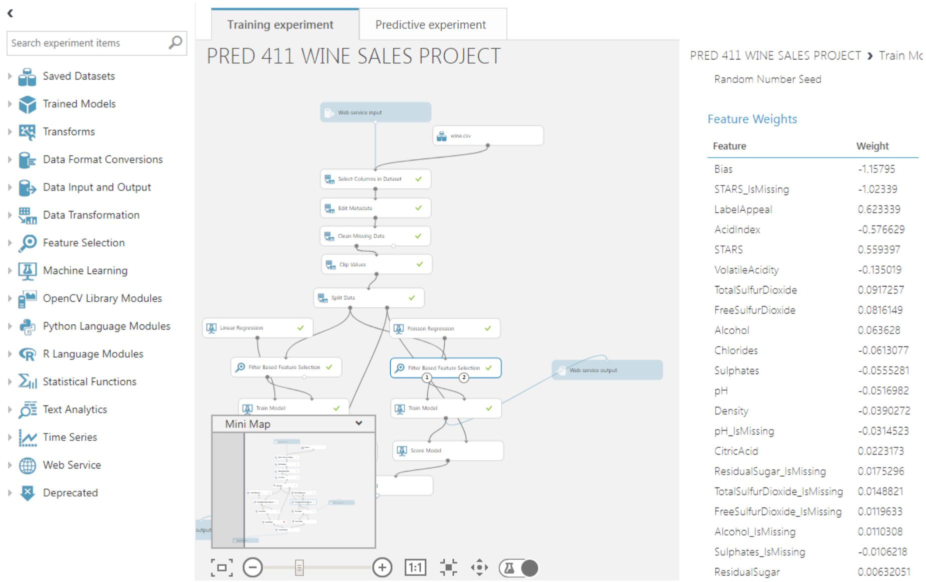

```
Effect                          DF     R-Square     F Value    p-Value

Var:   IMP_STARS                 1     0.153683   1626.145108   <.0001
Var:   LabelAppeal               1     0.068894    793.490408   <.0001
Var:   AcidIndex                 1     0.052548    649.021979   <.0001
Var:   VolatileAcidity           1     0.003873     48.090092   <.0001
```



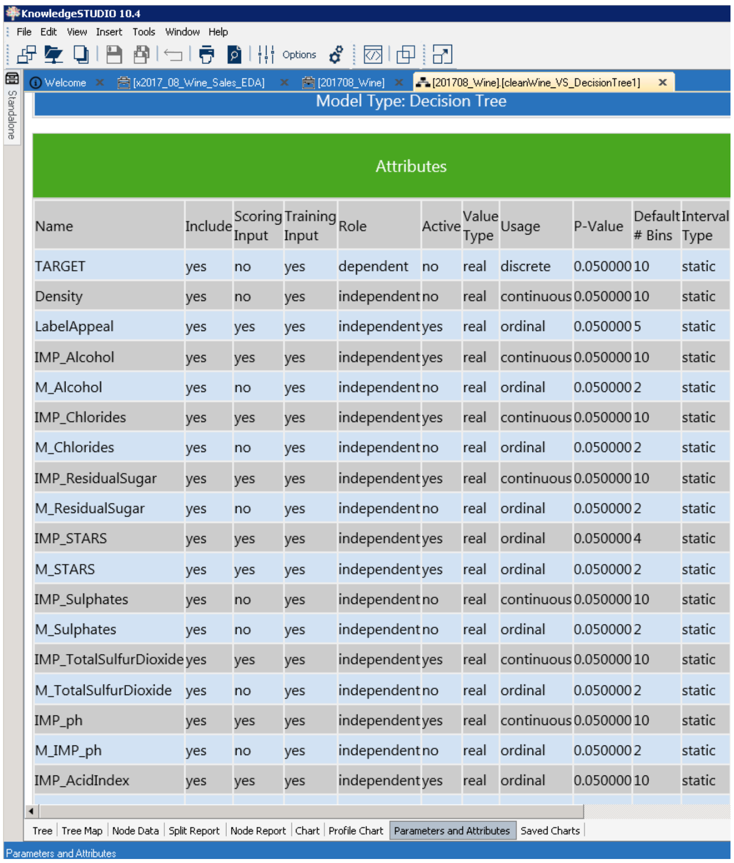
Scott Morgan
**KAGGLE:** ScottMorgan

## Microsoft Azure ML Poisson Feature Selection



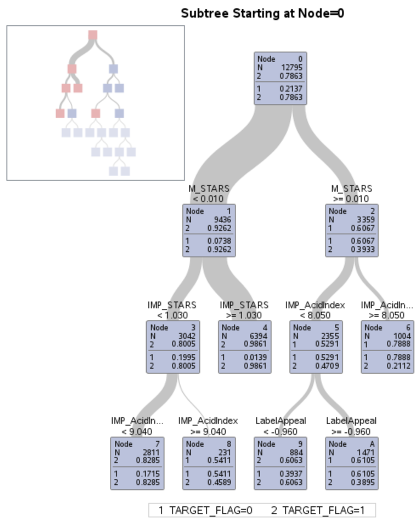

| Feature | Weight |
|---|---|
| Bias | -1.15795 |
| STARS_IsMissing | -1.02339 |
| LabelAppeal | 0.623339 |
| AcidIndex | -0.576629 |
| STARS | 0.559397 |
| VolatileAcidity | -0.135019 |
| TotalSulfurDioxide | 0.0917257 |
| FreeSulfurDioxide | 0.0816149 |
| Alcohol | 0.063628 |
| Chlorides | -0.0613077 |
| Sulphates | -0.0555281 |
| pH | -0.0516982 |
| Density | -0.0390272 |
| pH_IsMissing | -0.0314523 |
| CitricAcid | 0.0223173 |
| ResidualSugar_IsMissing | 0.0175296 |
| TotalSulfurDioxide_IsMissing | 0.0148821 |
| FreeSulfurDioxide_IsMissing | 0.0119633 |
| Alcohol_IsMissing | 0.0110308 |
| Sulphates_IsMissing | -0.0106218 |
| ResidualSugar | 0.00632051 |

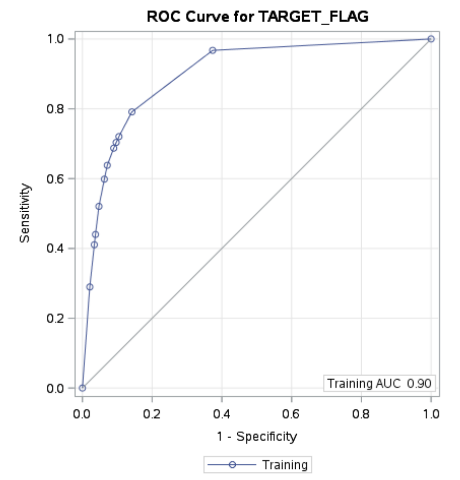
## Angoss Decision Tree




| Name | Include | Scoring Input | Training Input | Role | Active | Value Type | Usage | P-Value | Default # Bins | Interval Type |
|---|---|---|---|---|---|---|---|---|---|---|
| TARGET | yes | no | yes | dependent | no | real | discrete | 0.050000 | 10 | static |
| Density | yes | no | yes | independent | no | real | continuous | 0.050000 | 10 | static |
| LabelAppeal | yes | yes | yes | independent | yes | real | ordinal | 0.050000 | 5 | static |
| IMP_Alcohol | yes | yes | yes | independent | yes | real | continuous | 0.050000 | 10 | static |
| M_Alcohol | yes | no | yes | independent | no | real | ordinal | 0.050000 | 2 | static |
| IMP_Chlorides | yes | yes | yes | independent | yes | real | continuous | 0.050000 | 10 | static |
| M_Chlorides | yes | no | yes | independent | no | real | ordinal | 0.050000 | 2 | static |
| IMP_ResidualSugar | yes | yes | yes | independent | yes | real | continuous | 0.050000 | 10 | static |
| M_ResidualSugar | yes | no | yes | independent | no | real | ordinal | 0.050000 | 2 | static |
| IMP_STARS | yes | yes | yes | independent | yes | real | ordinal | 0.050000 | 4 | static |
| M_STARS | yes | yes | yes | independent | yes | real | ordinal | 0.050000 | 2 | static |
| IMP_Sulphates | yes | no | yes | independent | no | real | continuous | 0.050000 | 10 | static |
| M_Sulphates | yes | no | yes | independent | no | real | ordinal | 0.050000 | 2 | static |
| IMP_TotalSulfurDioxide | yes | yes | yes | independent | yes | real | continuous | 0.050000 | 10 | static |
| M_TotalSulfurDioxide | yes | no | yes | independent | no | real | ordinal | 0.050000 | 2 | static |
| IMP_ph | yes | yes | yes | independent | yes | real | continuous | 0.050000 | 10 | static |
| M_IMP_ph | yes | no | yes | independent | no | real | ordinal | 0.050000 | 2 | static |
| IMP_AcidIndex | yes | yes | yes | independent | yes | real | ordinal | 0.050000 | 10 | static |

**BINGO BONUS #6 <10 points>:** Used PROC HPSPLIT for TARGET_FLAG selection for zero model variable selection.

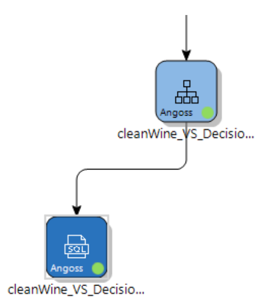


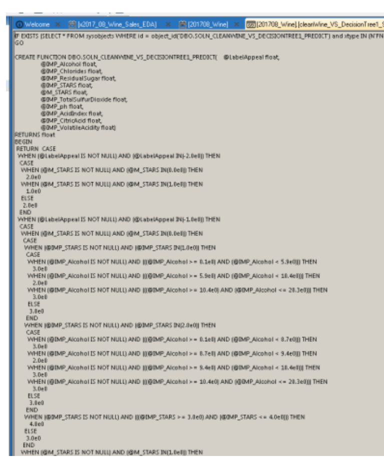



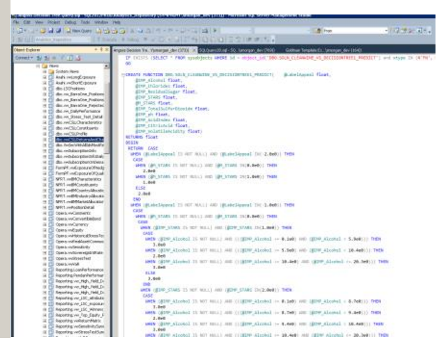

| Variable Importance | | | |
|---|---|---|---|
| | Training | | |
| Variable | Relative | Importance | Count |
| M_STARS | 1.0000 | 37.5141 | 1 |
| IMP_AcidIndex | 0.3424 | 12.8430 | 3 |
| IMP_STARS | 0.3177 | 11.9183 | 1 |
| LabelAppeal | 0.2170 | 8.1394 | 2 |
| IMP_VolatileAcidity | 0.1268 | 4.7563 | 3 |
| IMP_TotalSulfurDioxide | 0.1037 | 3.8898 | 1 |

**BINGO BONUS #7 <10 points>: Created SQL Function for Angoss Decision Tree**

I use SQL a ton every day for my job so I thought it would be interesting to create a decision tree function for this exercise. Very cool to know this can be done.

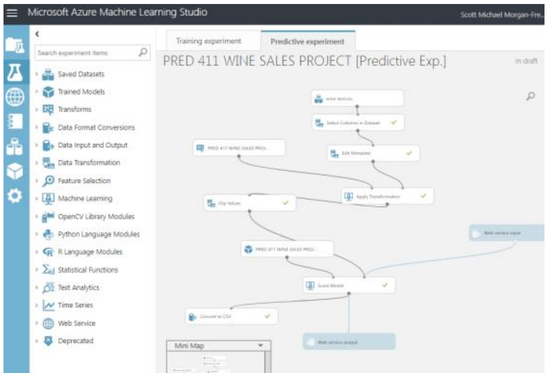


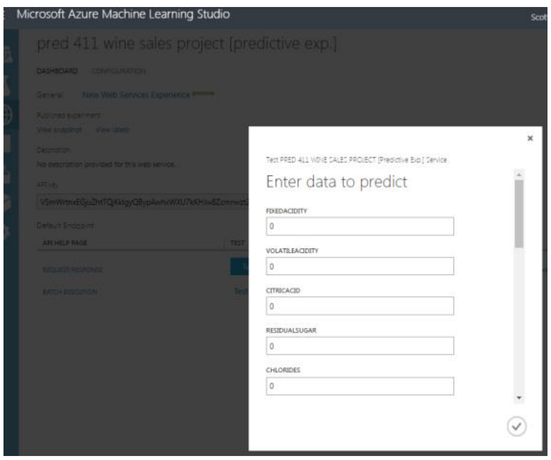





**BINGO BONUS #7 <20 points>: Setting up a Web based service for a model in Microsoft Azure ML**

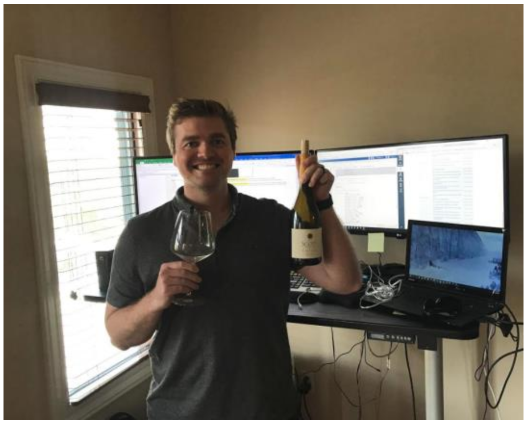
Created a web based interface for a trained Azure model / Predictive Experiment. You enter the values and it provides the predicted TARGET. Definitely some exciting applications for this type of user interface in my current and future roles.  I was impressed overall with how easy Azure was to use.




**BINGO BONUS #8 <10 points>:** You'll notice that I used SAS Macros in my file titled Homework_03_Scott_Morgan_ANALYSIS_Code.sas. I even included the periods to look like a pro.

**BINGO BONUS #9 <1000 points??>:** Sitting back and enjoying a glass of wine after finishing this paper. Opus One was a little expensive so I made my decision based on the label appeal (clearly). Turns out the models are right! The wine is actually pretty good too; first time trying Scott Family Estate Pinot Noir. Stick with the Arroyo Seco Monetery over the Russian River.



## REFERENCE(S)

Godden, P.W., Francis, I.L., Field, J.B.F., Gishen, M., Coulter, A.D., Valente, P.J., Hoj, P.B. and Robinson, E.M.C. (2002) An evaluation of the technical performance of wine bottle closures. Proceedings of the 11th Wine Industry Technical Conference. Eds. R. Blair, P. Williams and P. Hoj (AWITC: Glen Osmond, South Australia) pp. 44-52.