COMP 3948 Predictive Analytics   
Assignment 2

Aiman Ismail, A01052971, Set U

# Purpose:

Using a given dataset on car insurance, I have created multiple models to predict whether a automobile has been in an accident or not using CLAIM\_FLAG column as the y (target) variable. Using these models in an iterative process, I will explain

1. my exploratory analysis of the data,
2. how I prepared the data,
3. built my models, and
4. how I selected the best model and the trailing 2nd and 3rd best performing models.

# Exploratory analysis of data

To understand the composition of the data, I examined the general statistical information of the dataset by using. This generalized information is summarized the number of entries, missing entries, and what types of values each column contained. Using this information, there were several columns (categories) that were problematic.

Firstly, there were multiple columns that needed to be formatted as numbers. Secondly, there were five columns that required imputation to make up for missing entries. Thirdly, some categories required value boundaries to remove values that do not make sense. As a specific example to the third problem, Car Age had a value of -3.

Most importantly, my exploration of the data examined the distrubtion of the categories that had missing entries. This crucial step allowed me to identify distinct values that occurred frequently enough to impact the distribution of a category’s values. For instance, many of these same problematic categories had zero as a entry or not a number as an entry in the hundreds or thousands of occurences. Identifying the number of occurences and what type of values occurred in such numbers made it easier to determine my methods of imputation for these categories.

Categories (Columns) with Missing Entries

|  |  |  |
| --- | --- | --- |
| Income | Age | YOJ |
| Home\_Val | Car\_Age |  |

These key issues are resolve in my preparation of data.

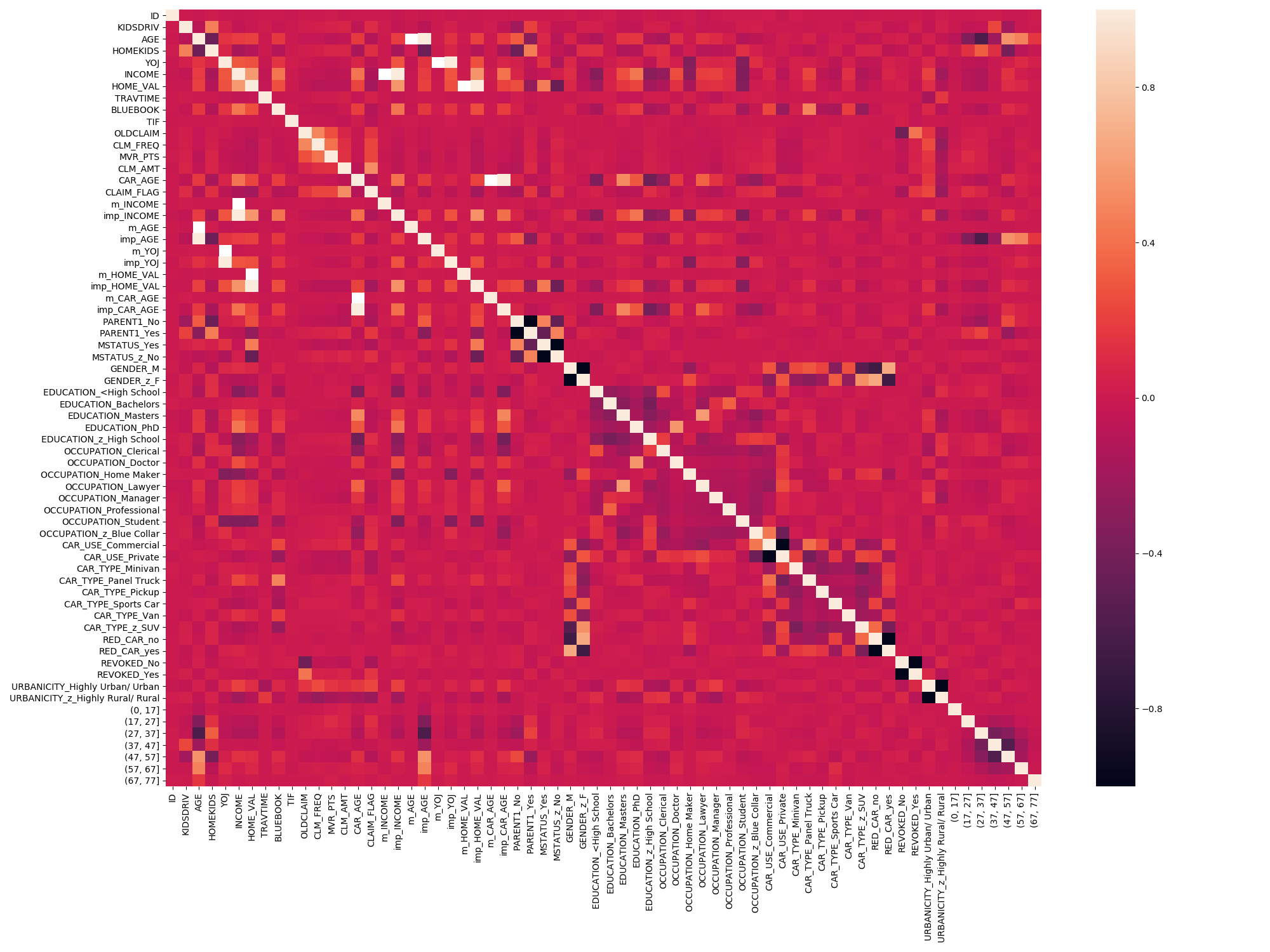
# PREPARATION OF DATA

# Selection of predictors

Our dataset has many points of information (columns). Having all this information was difficult to shift through to find meaningful predictor variables to determine whether a car was in an accident. To overcome this problem, I used multiple correlation heat map to narrow down likely candidates that have a highly negative or positive correlation with CLAIM\_FLAG column being 0 or 1 (where 0 is no accident and 1 is an accident occurred).

[Figure 1](#_Exploratory_Data_Analysis) shows a heat map for Model 1 with dummy variables and imputed variables introduced. Visually, we can see how the horizontal row of SalesPrice contain several indicators that influence price significantly. These predicator variables make up the basis of model 4 and help determine a predicted SalesPrice at given predicator variable levels.

Figure 1, Model 1, Heat Map



Using the correlation map, I ran regression on the likely predictor variables denoted by light orange hues and very dark purple hues. Below is a summary model 4’s linear regresion analysis.

|  |
| --- |
| Table 1 |

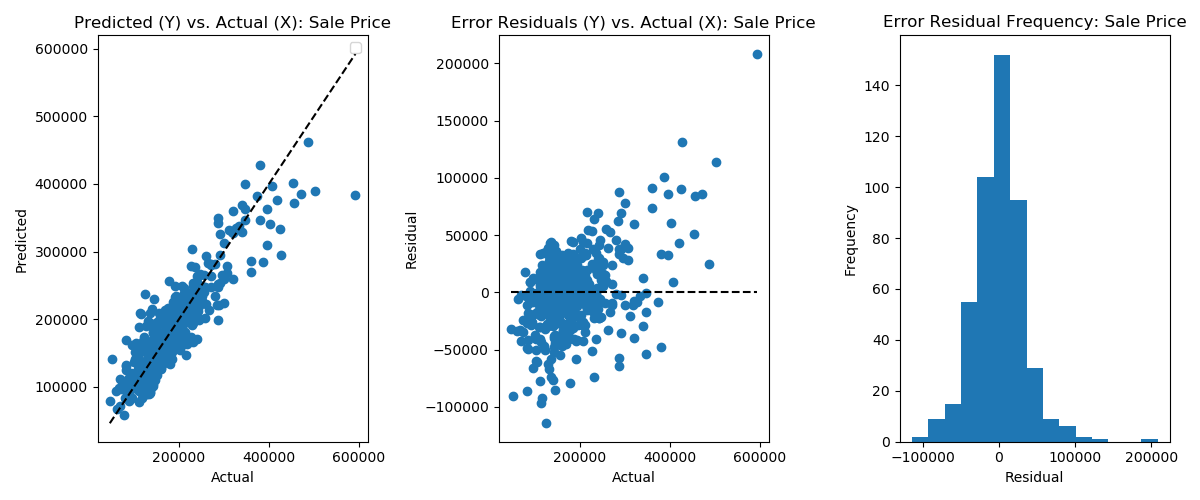
Refering the Table 1, we can see that aside from const, each of this predictor variables are contributing heavily towards SalesPrice since the majority of coef scores are in the tens of thousands range. Additionally, despite the coefs of LotArea, FirstFlrSf, and SecondFlrSf being less than 100 each in value these predictors have predictor values in the thousands (of square feet) and contribute the most to price.

Elaborating further on the predictor variables, it is the case that nominal variables of KitchenQual, ExterQual, and Neighborhood reflect whether a home has this nominal quality. This means that the coefficient contributes only if the predictor variable is present and contributes only once to the predicted SalesPrice. Despite this limitation, each category mentioned is adding or subtracting a range of $13580 to $52050 from SalesPrice. This is significant and is so much so that Kitchen Quality, Exterior Quality, and Neighborhood range in contribution from $27750 to $52050. The biggest detractor from price was ExteriorQuality being of a TA category. This removed $35370 from the sales price.

To get to Table 1, I ran linear regression iteratively, removing predictors with P>|t| scores that were greater than 0.05 level of significane. This occurred in Model 1 and 2. Model 3 involved removing predictors with P>|t| greater than 0. Finally, Model 4 involved removing predictors with coefficients that were not either very small or the resulting contribution to SalesPrice was small. For example a coefficient of 50 with a predictor value of 1000 would be considered small for my purposes.

Referencing Table 2, we can we see that Model 4 is very good at predicting sales price from the low-end up to around $350000. This can be intrepreted from the Predicted(Y) vs Actual(X): Sale Price Plot. Visually, we can see that the data points are more closely clustered around the middle to left portion. This indicates that our model requires more data on higher SalesPrice sales that exceed $350000. Another point to note is that the Error Residuals are closely clustered over the range of $44950 to $350,000. This is also where most data points occur. This suggests that our model has a high degree of explainable error. Later we will revist this explainable error by examining our Linear Regression Summary.

Table 2



# Model Evaluation

## Analysis of Models

Looking at how our models compared against each other, we selected several statistical measures to help compare which models perform well in specific categories.

This is a quick overview of the statistics being used for comparison.

|  |  |
| --- | --- |
| R2 | Coefficient of determination is a proportion of the variance in the dependent variables that can be explained by independent variables. Higher is better and the range is 0 to 1. |
| F-Statistic | A proportion that indicates the ratio of variable variance of SalePrice that can be explained by our model relative to variances due to error. Higher is better. |
| AIC | Measure of how over-fitted our model is. Lower is better. |
| BIC | Similar to AIC in evaluating over-fitted models. Penalizes more than AIC if there are more predictor variables. Lower is better. |
| Durbin-Watson | Determine whether error residuals are indepdent of each other. 0 to 2 shows positive correlation between residuals. A value above 2 but less than 3 indicates error residuals are indepdent (which is ideal). |
| RMSE | Average dispersion between actual and predicated values. Lower is better. (RMSE trends higher the less variables there are.) |

Table 3 – Summay of Models

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | # of Predictors | R2 | Adj. R2 | F-Statistic | AIC | BIC | Durbin-Watson | RMSE |
| 1 | 40 | 0.853 | 0.850 | 272.4 | 45230 | 45460 | 2.005 | 29756.443 |
| 2 | 22 | 0.850 | 0.848 | 487.7 | 45240 | 45360 | 2.010 | 29769.277 |
| 3 | 15 | 0.840 | 0.839 | 665.6 | 45340 | 45430 | 2.010 | 30721.030 |
| 4 | 10 | 0.812 | 0.811 | 826.2 | 45640 | 45700 | 2.045 | 33523.635 |

Of our models, Model 4 is the simpliest of the four. Simple means that it has the least predictor variables. Models 1 and 2 perform the best in R2 and adjusted R2 values. The differences across models is in the single percent deltas. This would mean that all four models perform quite well. Model 4 has the highest F-Statistic this implies that model 4 is roughly 3 times better than model 1 and 1.5 times better than model 2 at explaining errors. The AIC and BIC scores are quite similar across models but Model 1 and Model 2 perform the best. It is interesting to note that the BIC scores for Model 1 and 2 beat out Model 4 despite BIC punishing larger predicator numbers. All models have Durbin-Watson scores that hover above 2. This implies that our errors are independent of each other and are not related. Finally, models 1 and 2 perform the best in the RMSE statistic. This means that the delta difference of predicated sales price between actual sales price is smaller and thus better.

Lastly, all our models contain a high level of collinearity (or have high multicollinearity). This is expected in our model because some predicator variables are correlated with other predicator variables. In our models we have similar categories like Exterior Quality predictors, Neighborhood predictors, Kitchen Quality predictors, and First Floor Square Footage and Second Floor Square Footage predictors. Since these predicators are comparing closely related things we have an explanation and a reason to accept this high level of collinearity which occurs in all four models.

## Model Selected

When looking at the plots for all models within the Figures section of the report we can not immediately notice any signifiance differences between models, so the selection of our model is based of statistical measures. Based on our comparison of models 1 through 4, I have decided that each statistic we used to compare the models against each other should not be weighed equally. For instance, the differences in R2 across models is insignificant and there are more important indicators of performance like number of variables, F-statistic, BIC, and RMSE.

I prefer Model 3 for several reasons. Firstly, it has the 2nd lowest number of predictors at a roughly 1/3 of Model 1 but performs similarily to model 1. Secondly, Model 3 ranks 2nd for F-statistic which is more than double Model 1. Thirdly, Model 3 has the second lowest BIC score. Finally, Model 3’s RMSE score is only 3.24% higher than Model 1’s RMSE but still performs excellently in the other categories.

I believe that having a smaller set of predictor variables is important in understanding what predicators contribute the most to the size of Sales Price and Model 3 can predict Sales Price and the errors well with only 15 variables. This persuaded me in not selecting Models 1 and 2.

## Model Interpretation

Since I have selected Model 3 as my preferred model, we can interpret the linear regression summary to create an equation for our model that determines Sales Price.

Equation

Looking at our equation we can see that nominal categories are either increasing or decreasing sales price of our model. Additionally, despite our equation looking complicated it is reasonably simple in that nominal categories of Neighborhood, KitchenQual, Foundation, and ExteriorQual can be True or False.

This is meaningful to our equation because most variables and their respective coefficients will multipled out by zero if it is not an exhibiting characteristic. For instance, a house being sold can only exist in one place so only one Neighborhood can contribute to the sales price and the remaining neighborhood factors will not be counted.

Aside from these factors, LotArea, YearRemodel, FirstFlrSF, SecondFlrSF, Fireplaces, and GarageCars will be multiplied by integer values from 0 and in excess of 1. This means that these contributing areas can become quite significant if their corresponding predicator values are large. This is the case with YearRemodel, FirstFlrSF, and SecondFlrSF.

I like our equation because the categories being used as predicators make sense as to why they would increase sales price. Certain neighborhoods influence price positively or negatively and similarly levels of quality of excellent (Ex).

# Table of Figures

Model 1, Heat Map

