FINAL TEST

Daniel Bejarano
December 21st, 2014

Problem 1: San Francisco Construction Permits EDA

1 Introduction

The purpose of Part 1 of this report is to provide the results from an exploratory data analysis (EDA) on building permits with the purpose of understanding the building construction work in the city of San Francisco. The questions that will be answered include:

1. How are construction projects distributed according to their size and cost?
2. How does the city categorize the permits they issue?
3. What is the profile of permits based on whether it is new construction or modifications?
4. What are the differences between *estimated* and *revised* cost estimates?

2 Dataset

The dataset "Permits 2014" contains information on 49,842 building permits filed in San Francisco from November of 2000 to October of 2014. There are 41 variables in the dataset, which are: *application number, form number, file date, status date, status, status code, expiration date, estimated cost, revised cost, existing use, existing units, proposed use, proposed units, plansets, x15 day hold, existing stories, proposed stories, accessor stories, voluntary soft story retrofit, number of pages, block, lot street number, street number SFX, AVS street name, AVS street SFX, unit, unit SFX, company first name, company last name, constractor phone, company name, street number, street, street suffic, city, state, zip code, contact name, contact phone, and description of construction work*. A sample of the data is provided in Table 1.1.

Table 1.1: Sample data from "Permits 2014"

	APPLICATION..	FORM_NUMBER	FILE_DATE	STATUS_DATE	STATUS	STATUS_CODE	EXPIRATION_DATE	ESTIMATED.COST	REVISED.COST	EXISTING.USE	EXISTING.UNITS	PROPOSED.USE
1	#200509193264	6	9/19/2005	1/6/2014	ISSUED	9	1/6/2015	125000.00	125000.00	OFFICE	0	

3 Analysis

To derive conclusions from the dataset, a series of analytical and graphical methods was used. The following section contains four of the most relevant plots obtained.

3.1 Plots from EDA

After obtaining data summaries that contained means, standard deviations, medians, number of samples, etc., it was observed that the distribution of both *estimated* and *revised* costs was normal in its logarithmic form. A logarithmic (base 10) histogram of *revised.cost* is shown in Figure 1.1. First, note that the group of permits that appear at around 0 are projects that involve no construction at all, and were filed to indicate a change of use in the building, or revisions to previously filed permits, among other things. Second, using

cost as a proxy for size we can see that most permits were filed for small projects, which, as we will confirm later in the report, are for renovation projects. This can be easily observed by looking at the peak of the bell curve occurs at $10,000, which means that around half of the projects cost less than that. Even on the upper half of the curve, there are few large projects, with the number of projects costing more than $1,000,000 is on the hundreds. The largest project (cost-wise speaking) has a cost of $150,000,000 (or 10^8.17).

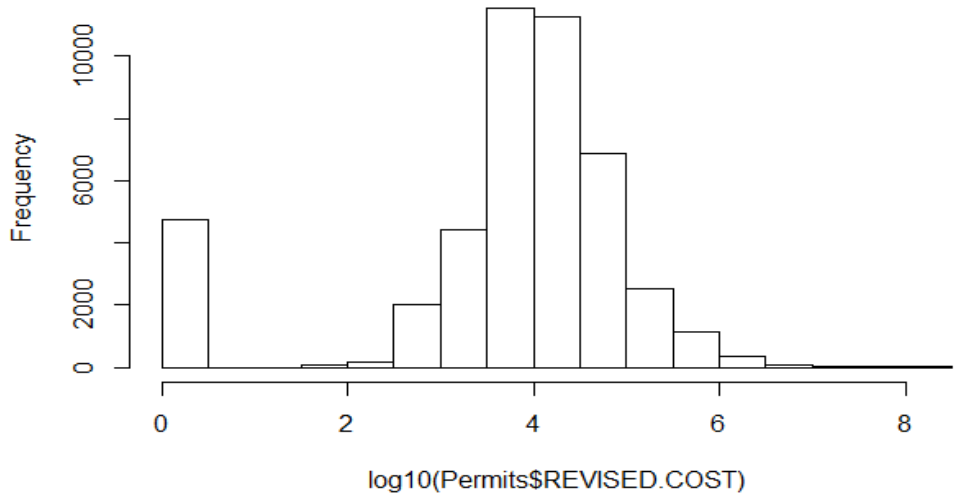

Figure 1.1: Histogram of Revised Cost from "Permits 2014"

To obtain a better understanding on the types of construction permits filed to the city, I looked at the *form numbers* they use to label each permit. As seen in Figure 1.2, some types of forms have a clear difference in their projects cost range. The most noticeable and statistically significant is One, but Two, Three, and Eight are also used to label permits that tend to be for projects that are more costly. An inference, which at this point can't be proven, is that One is used for projects that involve greenfield construction, as opposed to renovations. A closer look (histogram not included) also shows that permits of category Eight are the most abundant, at around 44,000. It can be inferred, although not concluded, that Eight is for mid-size renovation/modification projects, as they are the most abundant, and by looking at random samples of Eight permits they tend to fall under that category.

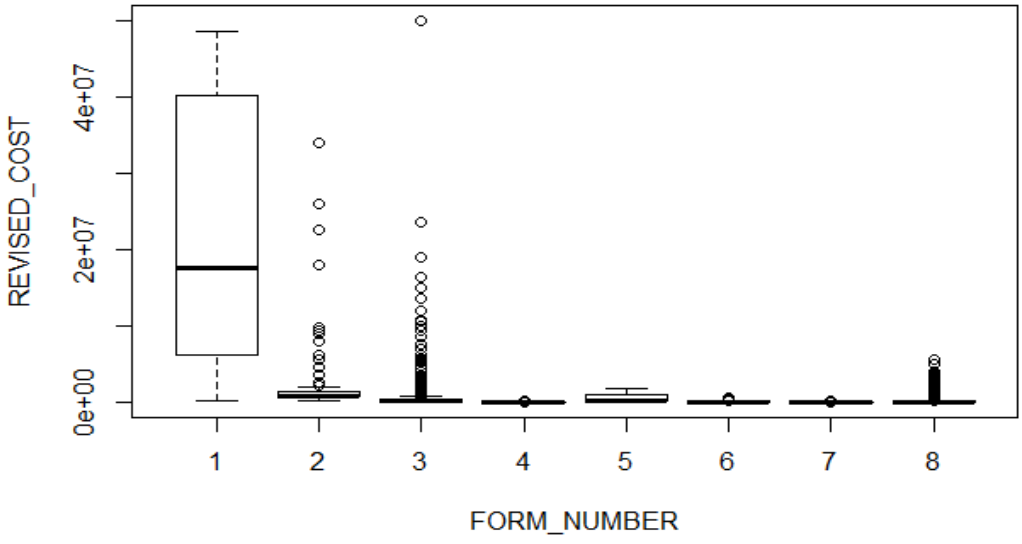

Figure 1.2: Revised Cost as a function of the categorical variable *Form Number*

To identify which construction projects involved new construction and which ones were modifications, Proposed Stories was plotted against Existing Stories (Figure 1.3). Although there are cases where new construction does not involve additional stories, this is still a good indicator for the overall dataset. The plot is color coded in the following way: **green** for projects under $100, **blue** for projects between $101 and $1000, **black** for construction costing between $1,001 and $10,000, and **red** for those above $10,000. A clear linear trend indicates that most of the project do not entail the addition of stories, especially in buildings larger than eight stories high. We can also observe that most of the red dots (due to resolution constrains not all of them are showing) are located on the lower left corner, and they tend to be characteristic of new projects, which explains to great extent the high cost. Additionally, green dots, are abundant in low-proposed and low-existing stories, a characteristic of small buildings. Since we had already inferred that projects under $100 were for the most part indication of change in the purpose of the building, with no additional construction, we can consequently infer that smaller buildings tend to mutate in their usage more than larger ones (which can be explained by the higher flexibility of smaller real estate). Finally, it seems like there is no project proposing the construction of a new large building, just as there is no project proposing the addition of stories to large buildings. Most projects that will add stories are either for new buildings, or very low ones (1 to 5 stories). There are also some few cases where demolition will occur, that is, *proposed stories* is lower than *existing stories.*

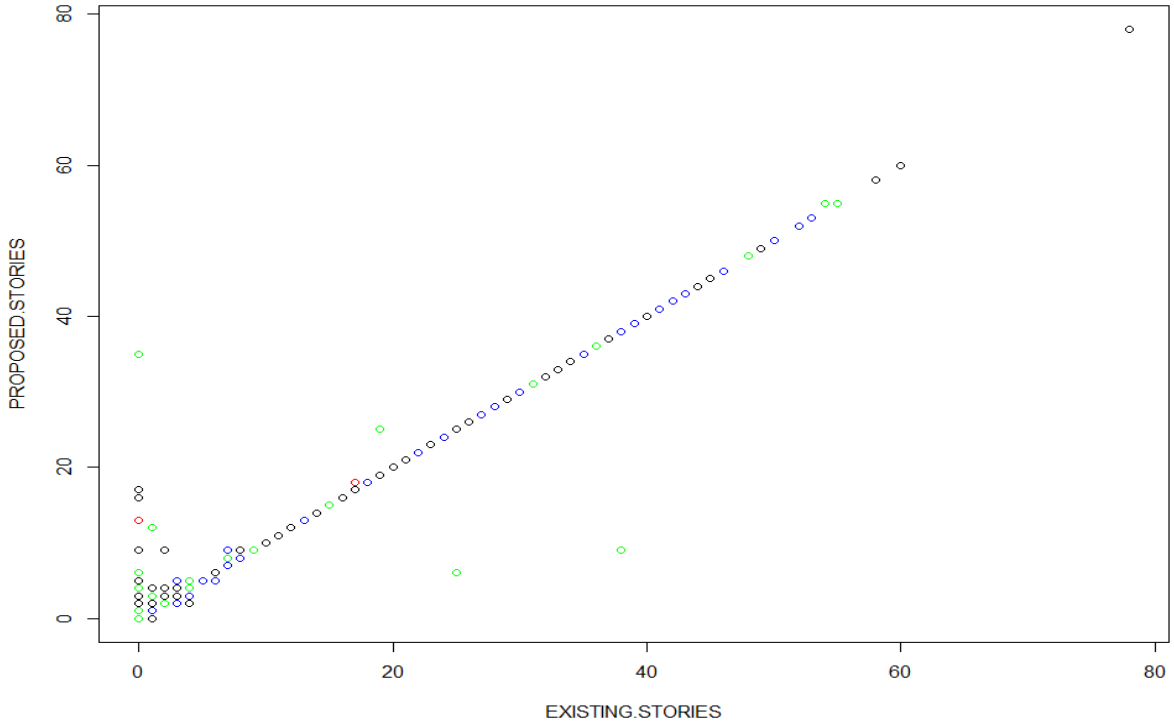

Figure 1.3: < 10^2 green, 10^2 to 10^4 blue, 10^4 to 10^6 black, >10^6 red

The last plot is Figure 1.4, and it shows on the y-axis the cost difference between *revised* and *estimated* costs (Diff_cost = revised.cost – estimated.cost). The linear, almost monotonic increasing, upper boundary is expected, as it signals that the difference between estimated and revised costs can at most be the actual revised cost (those cases where the estimated was $1.00, for instance).

An interesting observation is that a large number of cost estimates are given as rounded numbers (look at the periodic vertical trends, some of which are marked by the black rectangles); the separation between them tends to increase as costs are larger. This is expected due to the nature of how cost estimates are obtained, which use approximations because of the high levels of uncertainty prior to construction. Note also that for low revised costs, the cost difference is negative in several cases. What this suggests is that overestimates tend to be done more often for smaller projects, which are brought down once the original estimates are revised.

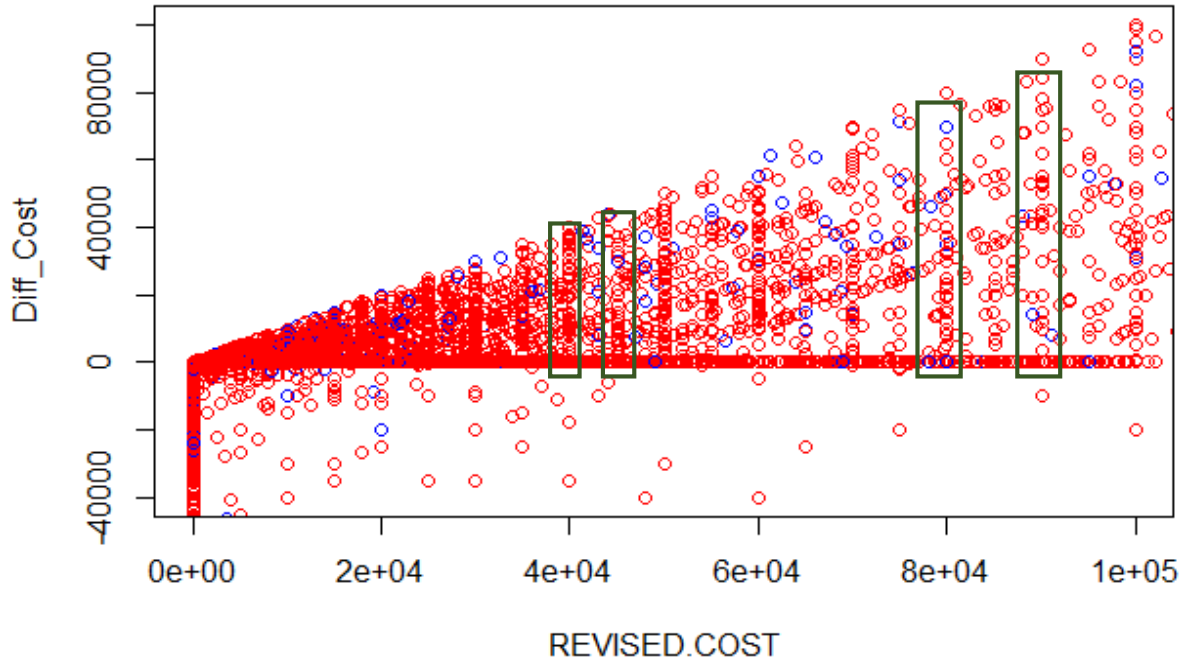

Figure 1.4: Cost Difference (estimated – revised costs) vs Revised Costs

3.2 Permit Fee Revenue for Affordable Housing

After exploring the data from a variety of different angles, it looked like categorizing the permits according to the proposed use of the building is prudent way of doing it. To do so, the 81 different proposed uses were identified, and then categorized as one of 10 categories: business, education, entertainment, food, hospital, housing, public services & utilities, telecommunications, and tourism. The reason for dividing in such categories is because these sectors encompass the most important ones an urban area requires to function properly. Therefore, each building can be charged a higher or lower fee as a means to incentivize certain types of development, or to charge more to those sectors that can, on average, pay higher fees. If, for instance, there is a lack of private sector involvement in the city, businesses could be encouraged to build by lowering or waiving their permit fees.

Table 1.2 contains the revenue that each of the categories would obtain if a 0.1% permit fee was charged. Each revenue is a function of the total amount of permits under that category (Business was by far the most populated) as well as the size of the projects under each category. Housing is the category with the largest projects, as can be observed by the revenue, $677,804. In addition, a more robust analysis was

performed to show the benefits of dividing categories in different tiers based on the cost of the projects. To achieve this, the distribution of the projects within each category was observed, which was used to make a separation between high and low tiers—where high means those projects that cost more, and low means those that cost less. An example is shown in the bottom part of Table 1.2, where the category of Business was divided and charged different fee percentages under the assumption that a business with more capital available would be more capable to pay higher fees, based on percentages. It can be observed that by charging 0.12% to the high tier and 0.08% to the low one, the total revenue was more than $8,000 higher than the original one (both highlighted by blue).

Table 1.2: Revenues from permit fees based on categories

Category	**Revenue**		
Business	$ 55,474		
Education	$ 6,398		
Entertainment	$ 5,514		
Food	$ 37,927		
Hospital	$ 3,164		
Housing	$ 677,804		
Public Services & Utilities	$ 709		
Telecommunications	$ 505		
Tourism	$ 45,481		
TOTAL	$ 832,976		
	Revenue	**%**	**Total Cost**
Business high	$ 58,141	0.0012	$ 48,451,000
Business low	$ 5,618	0.0008	$ 7,023,000
Business Total	**$ 63,760**		$ 55,474,000

4 Conclusions

One of the most challenging aspects of this problem was manipulating the data, as there were an abundance of entries lacking, and some of the formats (such as date) had to be converted. By performing EDA I was able to reach some preliminary conclusions that, although not conclusive, provide rich and important information about construction projects in the city. Via this methodology, it was identified that permits could be divided in categories that could be used to charge different permit fees. Although I did not explore this, it would have been interesting to see how long it took on average for projects to go from the filing of the permit, to its issuance and then see if they could be charged a higher fee when there was a project delay.

Problem 2: Building Temperature Model

1 Introduction

The purpose of Part 2 of this report is to describe the analysis performed to estimate the core temperature of a building via the usage of six sensors. To do so, different models of various complexities were tried. In the report, the following questions will be answered:

1. What is the distribution/profile of the sensors measurements data and temperature?
2. What is the expected performance of the final model?
3. What can be said about the confidence intervals for individual estimates of core temperature?
4. If the uncertainty in the temperature estimate similar across the entire range of temperatures?
5. If only two sensors could be used, which ones should be chosen?
6. How could a model written by another analyst, who claims its errors are what one should expect, be validated?

2 Dataset

The dataset includes 120 readings from six different sensors in a building, labeled x1, x2, x3, x4, x5 and x6, as well as the temperature readings on the same building, y. The sensors output is in voltages, and the temperature readings in Fahrenheit. A sample of the data is shown in Table 2.1

Table 2.1: Sample data from "Temperatures"

	x1	x2	x3	x4	x5	x6	y
1	0.91385464	0.68456784	1.80714798	2.56492905	0.288025867	2.89577392	61.53407
2	3.61364099	0.29900566	3.72131528	1.57638640	0.044363676	2.33444138	70.54751

3 Analysis

Before developing a model, the data output from each sensor was graphed as a histogram to obtain a better understanding on its underlying distribution. It was observed that none of them had a normal distribution, but that "y" approximated one when its log was plotted instead. A plot of each variable against each other (pairs plot in Figure 2.1) showed a very low correlation amongst the explanatory variables. It can be seen, though, that there is some relationship between the explanatory variables and the temperature readings. From the plot a nonlinear relationship can be observed, which will serve as the basis for developing the model.

Throughout this section, Model 1, Model 2 and Model 3 will be used to refer to the three best models developed in the categories of: (1) linear in parameters and variables, (2) linear in parameters but non-linear in predictor variables, and (3) non-linear. Model 4 refers to the polynomial model, and Model 5 refers to the MARS model, which was the best one obtained. Some other more complex nonlinear models were tried, but provided inferior results when compared to MARS. The characteristics of each are explained next:

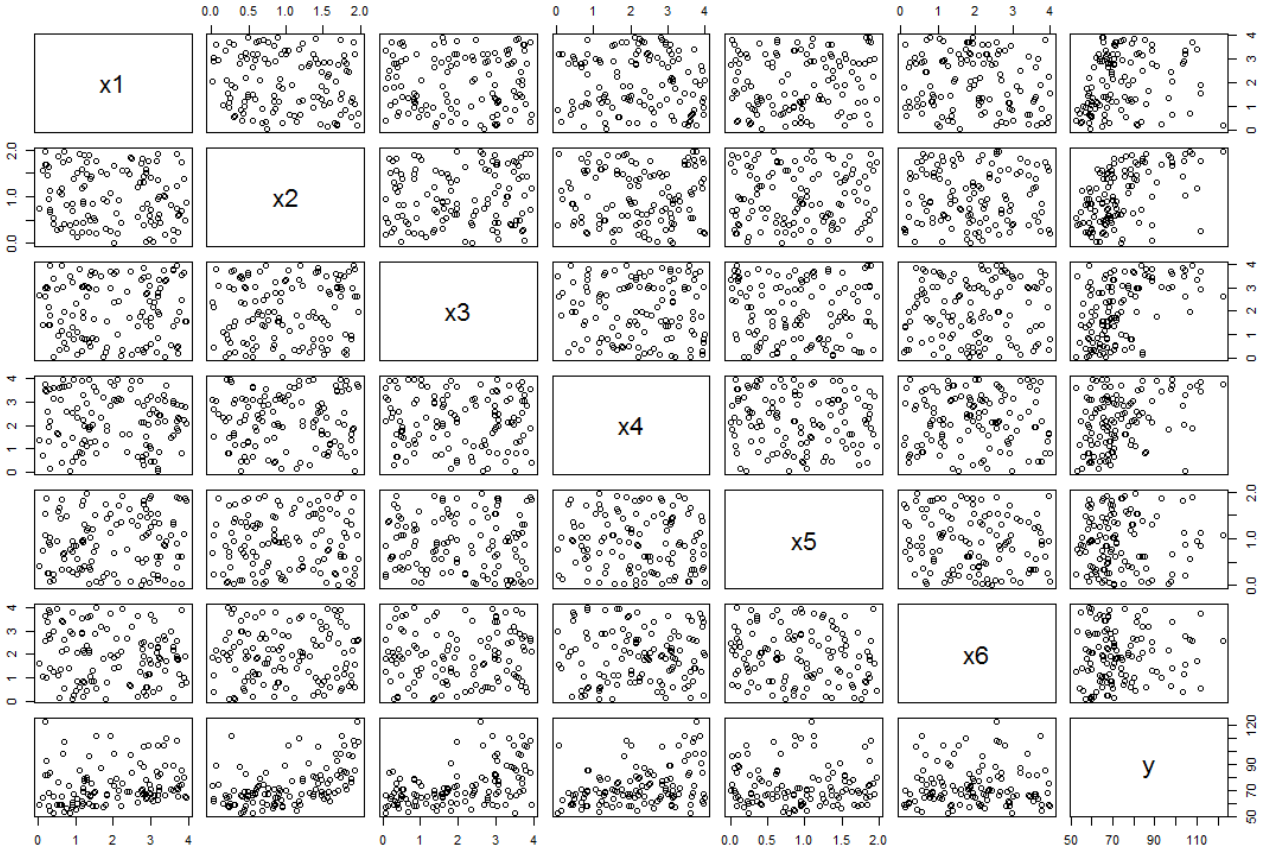

Figure 2.1: Histogram of Revised Cost from "Permits 2014"

3.1 Development of Models to Output Core Temperatures

The first model tried (Model 1) was linear regression using forward variable selection. The best model using this method was to include variables 1 through 4, as shown in Figure 2.2. The adjusted R-squared was a rather low 0.6649, and by looking at the residuals plot (section 3.3) it is evident that a nonlinear model would perform better.

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = temp)

Residuals:
    Min      1Q  Median      3Q     Max
-21.937  -5.635  -1.063   5.127  31.907

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.7672     2.9935  10.278  < 2e-16 ***
x1            3.7846     0.6826   5.544 1.90e-07 ***
x2           13.2129     1.3870   9.526 3.29e-16 ***
x3            5.3899     0.6519   8.269 2.68e-13 ***
x4            5.0395     0.7060   7.138 9.13e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.583 on 115 degrees of freedom
Multiple R-squared:  0.6762, Adjusted R-squared:  0.6649
F-statistic: 60.04 on 4 and 115 DF,  p-value: < 2.2e-16
```

Figure 2.2: Summary of Model 1

Model 2 was created as linear in its parameters but making the predictor variables nonlinear. The methodology was also a forward variable selection, where variables would be added one at a time but always trying a set of transformations, mainly: log, exp, inverse and different powers. The summary of the model and its performance is shown in Figure 2.3. Adjusted R-squared was improved to 0.7541, and one additional variable, x5t, was included. The relationships between the original variables and the transformed ones (x1t, x2t, etc.) is as follows:

```
Call:
lm(formula = y ~ x1t + x2t + x3t + x4t + x5t)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2687  -4.9746  -0.6794   4.7096  23.7269

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.07454    1.75332  27.419  < 2e-16 ***
x1t          1.86818    0.28139   6.639 1.12e-09 ***
x2t          0.16377    0.01306  12.541  < 2e-16 ***
x3t          1.30955    0.13929   9.401 6.90e-16 ***
x4t          2.27899    0.28273   8.061 8.33e-13 ***
x5t         -0.38202    0.13563  -2.817  0.00572 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.353 on 114 degrees of freedom
Multiple R-squared:  0.7644, Adjusted R-squared:  0.7541
F-statistic: 73.97 on 5 and 114 DF,  p-value: < 2.2e-16
```

Figure 2.3: Summary of Model 2

```
x1t = (temp$x1)^1.5
x2t = (temp$x2)^5
x3t = (temp$x3)^2
x4t = (temp$x4)^1.5
x5t = log(temp$x5)^2
x6t = (temp$x6)
```

By looking at the residuals it is still evident that a nonlinear model could account for part of the error that cannot be explained by randomness.

Although other models were tried, including a better MARS model, Model 3 will be used to answer the questions on this part of the problem, as I manually created this model and understand it better than MARS. Model 3 was created as a multiplication of the previous transformed variables (in this case the model improved when adding x6t, so x1t through x6t). The resulting adjusted R-squared improved to 0.8589, and the residuals plot also indicate a better performance, as shown in Figure 2.4. Meaning, there is no clear pattern by looking at the plot, and the errors bounce randomly around zero, thus indicating a good fit. However, it can be observed that the predicting power of the model is greater for temperature

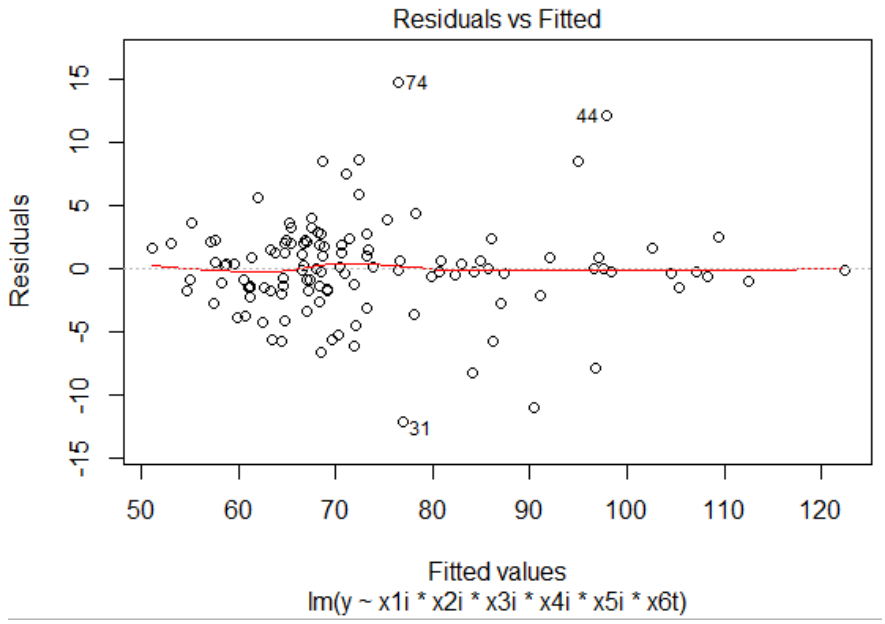

Figure 2.4: Model 3 Residuals vs Fitted Plot

values lower than 80 degrees. The range between 80 and 100 degrees has a tendency towards negative numbers, which should not be obtained from random error. Therefore, the validity of the model decreases in such range due to a higher level of uncertainty.

Both the p-value and the standard error are low, which indicates that the model represents the data in a statistically significant way. As for the confidence intervals, even though the errors are low, the model has been developed to fit the data provided, and while the best form of determining how confident the predictions are would be to test it on a test dataset, we used the entire dataset to train. Thus, the results would be highly biased and indicate a lower error than in reality it would with a dataset the model has never seen before. Using 2.5% and 97.5% confidence intervals, the values obtained from Model 3 were: -0.08 and -0.04, respectively.

Other models tried were MARS, which chose a degree of 2 and fit parameter of 14 as its optimal result, and provided an adjusted R-squared of 0.899, a slight improvement from Model 3. Another method tried was polynomial regression, which served as a confirmation that variables x1 through x5 were the most significant ones (since the algorithm has the capability of selecting them automatically), something that had been observed in Model 1 and 2. A last model used was Hyfis, which provided results that were not as good as those obtained from the previous models.

3.2 Best Model Formed by Only Two Explanatory Variables

Although there are several methods that as part of their algorithm they perform variable selection, the model formed by only two variables was created manually by forward variable selection using linear regression but doing variable transformation.

The methodology used was to look at which combination of variables provided the highest adjusted R-squared value, and presented a residuals plot that complied with the criteria that has been mentioned in section 3.1 and will be further explained in section 3.3. The final 2-variable model was: **y = x1t + x2t.** The adj R^2 is 0.4388, which is much higher than any of the other combinations by an order of magnitude in most cases. The residuals plot, Figure 2.5 left plot, show a negative tendency, and are grouped mostly in the lower temperature values. It is evident that points don't bounce randomly off zero, but that is the case for most of the models that can be generated with only two variables—such as the right plot in Figure 2.5, which is **y = x1t + x6.**

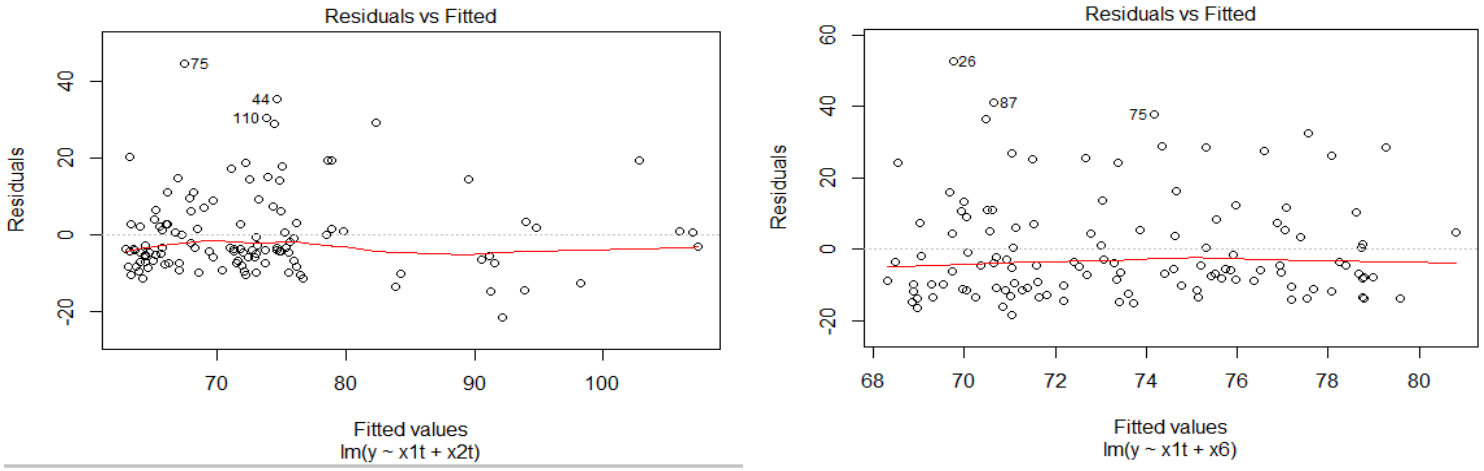

Figure 2.5: Residuals vs Fitted comparison between two two-variable regression models

3.3 Validation of other Programmer's Model

With the information provided, it can't be known for certain whether the model by the analyst is "good" or "bad". Regardless, what he claims is correct: errors are expected when one creates a model to represent events in real life. The reason is that randomness and unpredictability are necessary components of models, which are composed of a deterministic and a stochastic component. The deterministic is what can be explained by the variables included in the model, while the stochastic part is the error attributed to the randomness of real-world events.

To validate his model, several tests can be performed. The first thing is to check the Adjusted R-squared and ensure it is a high value (preferable compared to the R-squared, as the former penalizes the addition of explanatory variables and reduces the apparent improvement on the model from just adding variables to it). This is not a conclusive test though. A more robust tests is the Residuals vs fitted plot, which can be used to analyze whether the error obtained from the model is consistent with random error, and therefore unpredictable. For this reason, the residuals should not be predictable, or have any structure or pattern. Figure 2.6 shows two plots from two different models, where the one on the right (Model 2) is a better model than the one on the left (Model 1). One can see how Model 1 residuals have a trend, which is marked by a red line. Model 2 has a less pronounced trend, indicating that more of the deterministic component of temperature can be explained by the Model 2, but not all of it. In this case, the conclusion is that the linear assumption is not valid and the data could be better explained by a nonlinear model.

Similarly, the model created by the analyst can be checked using the same methods. Other things to look for are the amount of explanatory variables in proportion to the amount of data points used to create the model—a good rule of thumb is to limit the number of variables to one-tenth of the data entries.

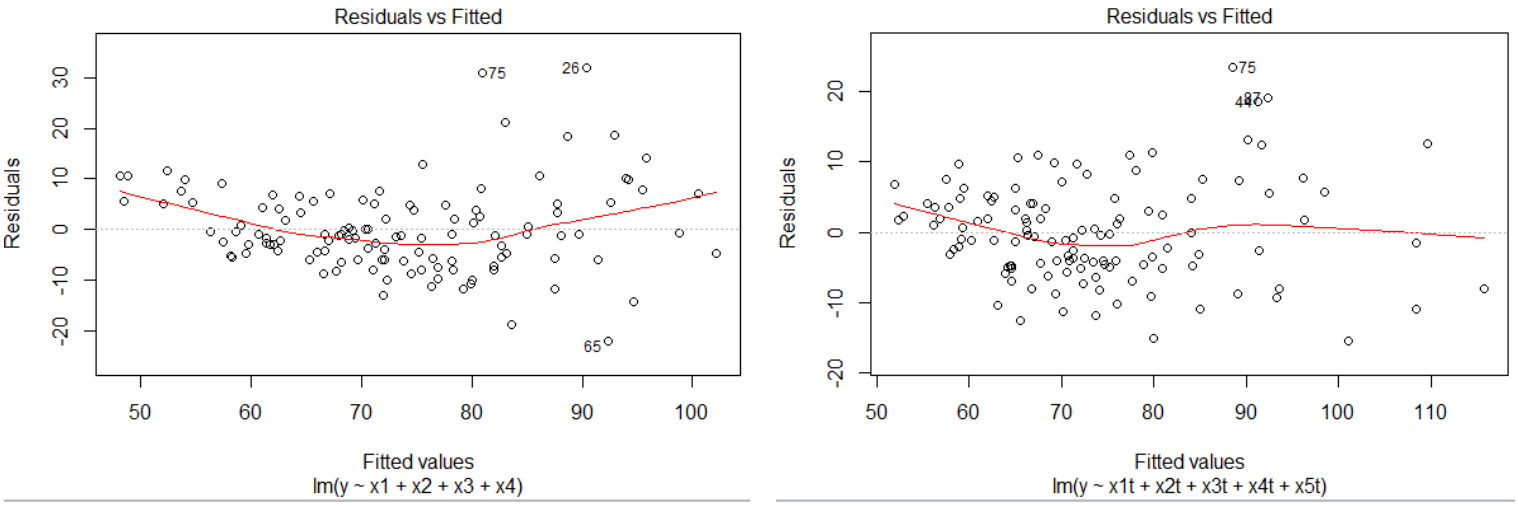

Figure 2.6: Comparison of Residuals vs Fitted for two different models

4 Conclusions

By tying different linear and nonlinear combinations of explanatory variables, a final model with a strong prediction power was developed. The strengths of higher-complexity models became evident both by their capacity to automatically evaluate a range of parameters and select optimal ones, and of course by their capacity to fit data that has a non-linear dependency and/or behavior.

Problem 3: Diabetes Predictive Model

1 Introduction

The purpose of Part 3 of this report is to show and analyze the results obtained from fitting a series of models that were then used to predict the probability of a person having diabetes, given a set of factors. By trying different methods and a combination of variables, the following questions were answered:

1. What is the performance obtained from doing a model without using variables that require "expensive" testing to obtain?
2. Based on the ROC curves, what are the performances of the different models tried?
3. How do performances compare for different methods, and for different models generated by combining variables in varying forms?
4. Which two of the variables in the models are most informative in making predictions?
5. What is the performance of models that include measurements from "expensive" testing?

2 Dataset

The dataset "Diabetes" contains information about diabetes in a certain population of 768 cases. The variables included are: *number of times pregnant, plasma glucose concentration at two hours in an oral glucose tolerance test (mg/dL), diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), two-hour serum insulin (mm U/ml), body mass index (weight in kg/(height in m)^2), diabetes pedigree function, age (years), class variable (0 – no diabetes, 1 – diabetes).* A sample of the data is shown in Table 3.1.

Table 3.1: Sample data from "Diabetes"

	preg	glucose	bp	triceps	insulin	bmi	pedigree	age	class
1	6	148	72	35	NA	33.6	0.627	50	Class1
2	1	85	66	29	NA	26.6	0.351	31	Class0

3 Analysis

All models were created using one of two methods: penalized logistic regression (LR), and support vector machine (SVM) with radial basis function kernel. For each analysis both methods were tried, and a comparison of the results obtained for each will be presented. It should be noted that while SVM is ideal for datasets with a large number of factors, it tends to not perform as great on datasets with very few attributes.

For all three parts of this section (3.1, 3.2, and 3.3) the data was split between train and test (80% and 20% respectively), and cross-validation was performed on the training data in order to fit and validate the model. All of this was done automatically by the algorithms in each of the two methods.

3.1 Analysis on Variables Not Requiring Lab Work and Filtering out NA Values

First, two models were created to predict individuals diabetes by including only those variables that did not require lab work (all but *glucose* and *insulin*). Additionally, only those data entries that had no NA values across all variables considered.

Combinations of different variables were tried to obtain the model that optimized the true positive rate. The determining criteria was to obtain the highest area under the ROC curves (*from now on this area will be referred to simply as ROC), but taking into consideration the number of variables that were going into the model, with some preference to having less instead of more variables included. A forward variable selection approach was taken, and to determine which variables to include I used the **regsubsets** function in R, which shows what variables are the most informative for any given number of variables. For instance, In Table 3.2 it can be seen that for one variable the most informative factor, and therefore the one that should be used, is *age.* It was found that this methodology resulted in models with the highest ROC for a given number of variables. Then, through trial and error the number of variables was determined by observing which resulted in the highest ROC.

```
Selection Algorithm: exhaustive
         preg bp  triceps bmi pedigree age
1  ( 1 ) " "  " " " "     " " " "      "*"
2  ( 1 ) " "  " " " "     "*" " "      "*"
3  ( 1 ) " "  " " " "     "*" "*"      "*"
4  ( 1 ) "*"  " " " "     "*" "*"      "*"
5  ( 1 ) "*"  " " "*"     "*" "*"      "*"
6  ( 1 ) "*"  "*" "*"     "*" "*"      "*"
```

Table 3.2: Variable Selection Results

The resulting LR and SVM models obtained were:

- **LR: class = bmi + pedigree + age,** and
- **SVM: class = bmi + pedigree + age + preg**

The corresponding coefficients and parameters will not be included in the body of this report, but can be easily obtained from the code. As mentioned earlier, it makes sense that SVM required more variables than LR, and is also interesting to see that even then SVM performance was inferior. This entails that the data may be more easily separated by linear boundaries given the variables used in this case.

After training the models, they were tested and their performance was evaluated based on their ROC curves, shown in Figure 3.1. The performance of the LR model (left graph) is superior to that of the SVM (right graph), with ROCs of 0.8064 and 0.7701, respectively. We can also observe that the LR curve has a steeper slope at high specificity values, which results in a higher sensitivity when compared to the SVM performance at most of the specificity range. The purple and blue lines mark a specificity value of 0.2 and a sensitivity value of 1.0, respectively. Notice that although LR performs better overall, both models have an almost identical sensitivity at the maximum false positive rate we are interested in, 0.2. This shows that model performance is relative to the criteria with which a model will be evaluated, and depending on the threshold models perform differently.

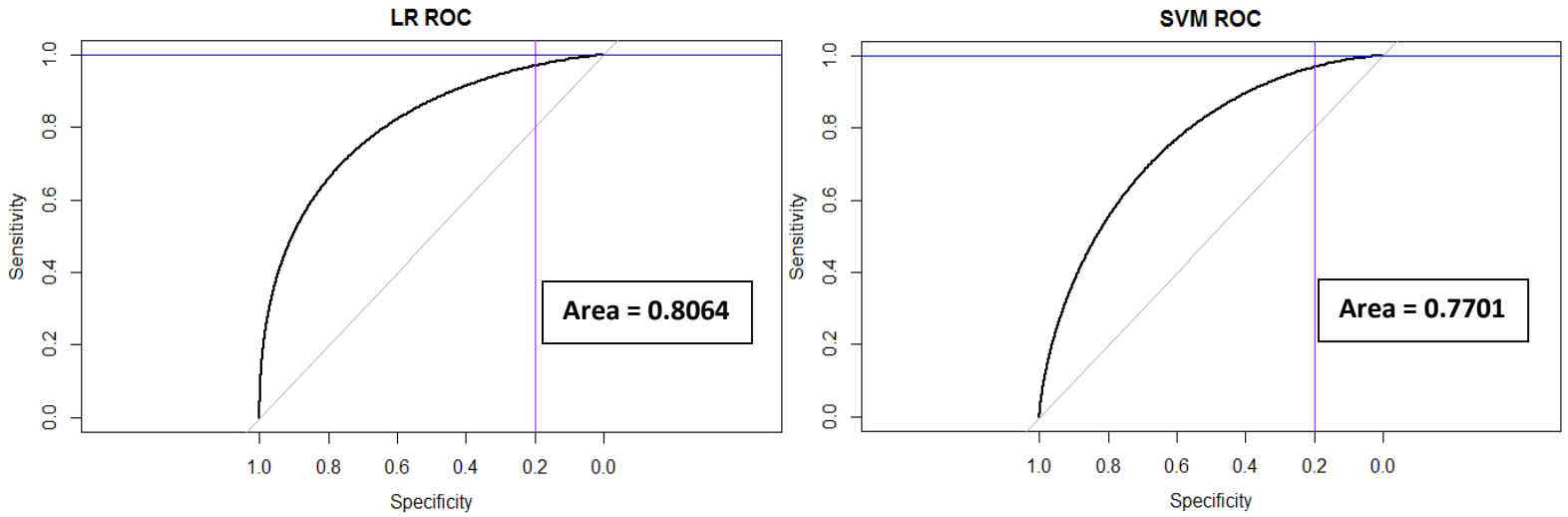

Figure 3.1: ROC Curves from LR and SVM "cheap testing" models

3.2 Two-Variable models with Most Informative Factors

As mentioned in the previous section, an R function was used to determine which variables are the most informative. This can also be determined by looking at the previous LR model and selecting those variables that are marked by "*" as the ones that are most relevant. As shown in Table 3.2, the two most informative variables when predicting diabetes, without considering *glucose* and *insulin,* were ***age*** and ***bmi.*** The ROC curves for these two models appear in Figure 3.2, and similar to the previous case, the simpler LR model outperforms the more complex SVM by resulting in a higher ROC (0.7852 vs 0.7577). In this case, LR has a higher True Positive (TP) Rate at our selected maximum False Positive (FP) Rate of 0.2. At this rate of non-diabetic individuals incorrectly classified as diabetic, and based on our date of course, more diabetic patients will be classified as Class 1 when using the LR model compared to the SVM model. Lower TP rates comes at the expense of having higher TN rates, which result in more individuals unnecessarily going through additional testing.

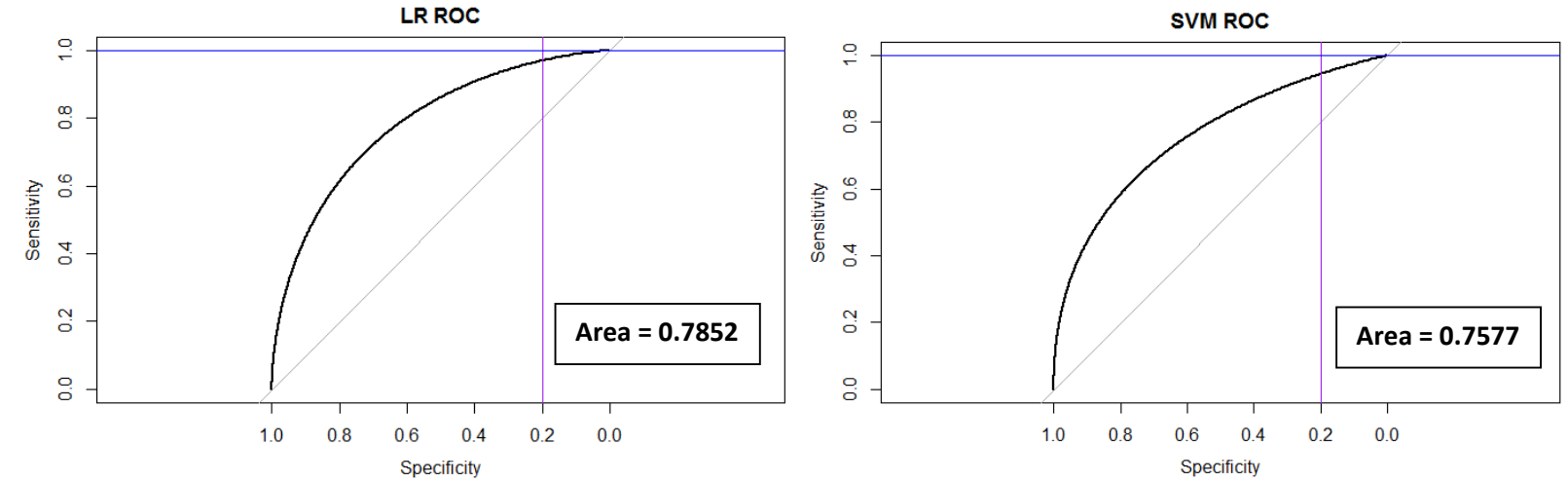

Figure 3.2: ROC Curves from LR and SVM two-variable "cheap testing" models

3.3 Model for Entire Dataset with Substituted NA Values

The last set of models were trained and tested on an engineered dataset, where the missing values (NA) in the original Diabetes dataset were substituted with predictions obtained via a linear regression model. This time, all of the original variables were included, totaling eight explanatory variables and *class* as the predicted one. The methodology for substituting the NA values was to train on all the data entries where no variable had a missing value. Then, a linear regression model was trained for each variable, and whenever a variable had a missing value it was predicted based on the values of the other variables. Whenever there were several NAs in the same row, only those factors without NAs were used, and once a prediction had been made on one factor it was then used to predict on the others. This proved to be a better method than just inputting the average value for each variable, which was proved after observing that the standard deviation (SD) of each variable was higher than the root-squared mean error (RSME) obtained from each regression. Table 3.3 contains the values as well as the percentage difference (with RMSE as the base) for each variable that had missing values. In some cases the difference was minimum, and a regression model may not have been necessary.

Table 3.3: SD and RMSE comparison for filling missing values

	SD	RMSE	Δ
Glucose	30.86	26.24	18%
Bp	12.49	11.89	5%
Triceps	10.51	10	5%
Insulin	118.84	97.06	22%
Bmi	7.03	5.05	39%

Having populated the dataset with the predicted values, the data was divided in train (80%) and test (20%). Using the same two methods as before, and the same methodology for variable selection, the best models obtained were the following, in both cases:

- **class = preg + glucose + bmi + pedigree**

The ROCs were 0.8541 and 0.807 for LR and SVM, respectively. For both models, adding a fifth variable resulted in a lower ROC. From the ROC curves in Figure 3.3 we can observe that: (1) the LR model outperforms the SVM one again, thus supporting the initial assumption that the classes are closer to be linearly separable than to exhibit a more complex behavior; (2) both models perform better than the previous cases, which allows us to conclude that the "expensive" methods provide values that are more informative; (3) the variables included in the model are similar to those used in the first couple of methods, but in this case *age* was dropped and *glucose* was included instead. In fact, *glucose* proved to be the most informative variable based on the dataset provided. Considering that a 0.85 TP rate was obtained with such a simple model and using factors that are not too complicated to obtain, I would qualify the models as being very powerful and an excellent tool as a preliminary diagnosis to evaluate the likelihood of an individual having diabetes.

I was surprised to see that insulin was not one of the variables chosen by the algorithm, so I did some exploratory data analysis and found that there is, as expected, a very high correlation between glucose and insulin, which is presented graphically in Figure 3.4. It then makes sense that one would be included but not the other, as having both would result in some redundancy in the model.

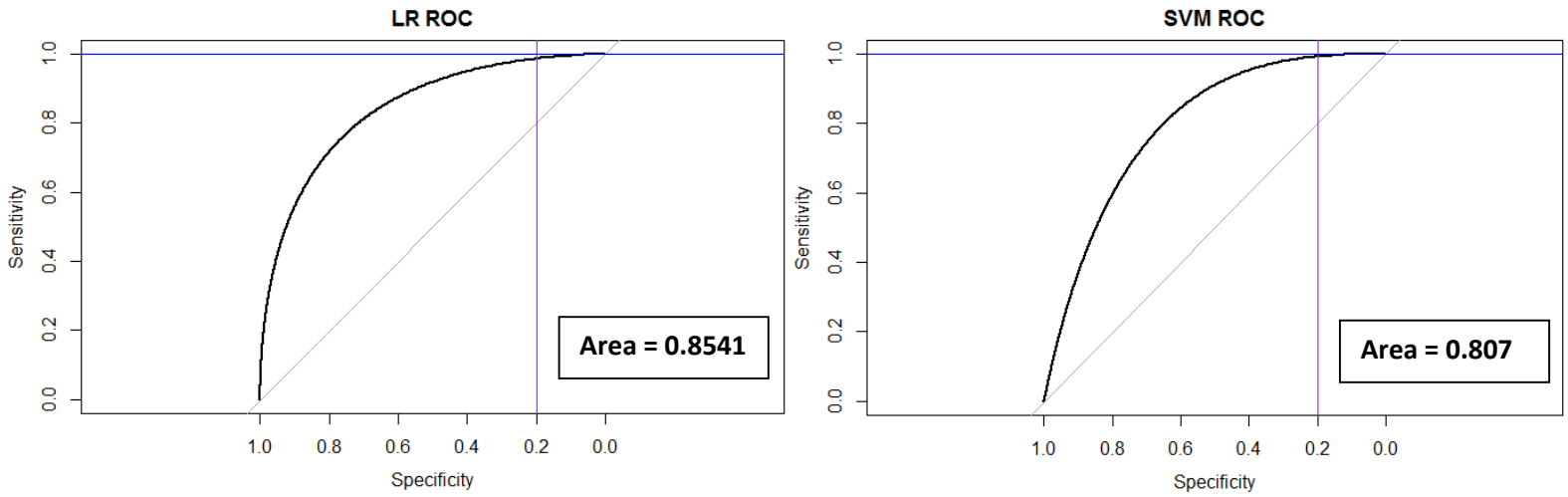

Figure 3.3: ROC curves for LR and SVM models based on entire dataset with substituted missing values

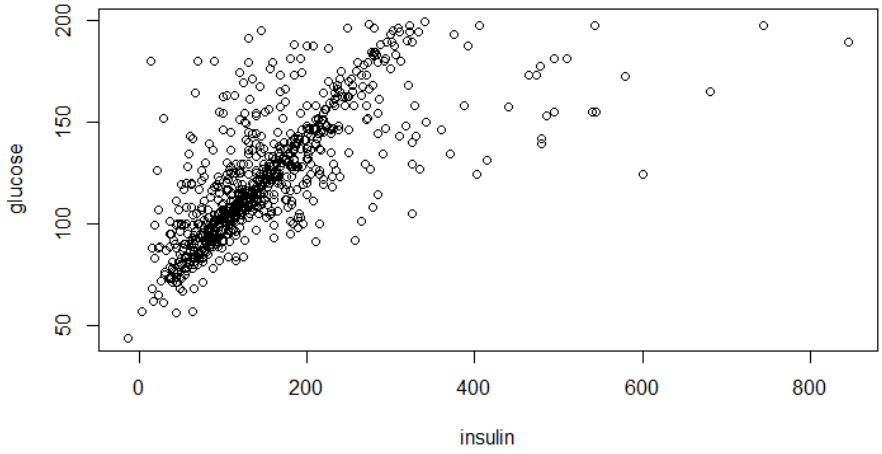

Figure 3.4: Plot of *glucose* vs *insulin*

4 Conclusions

The most important conclusions from Part 3 of this report were:

1. At an ROC of 0.8, a decent model can be developed using factors that can be easily and cheaply obtained. Even after including the more expensive tests, such as glucose levels, the ROC only increased to 0.85.
2. The LR model proved to have better results in predicting diabetes compared to the SVM model. This shows that higher complexity does not entail better results, and signals that certain models are better suited for different purposes; in this case we could observe that SVM performs better when more variables are available, and is better suited for variables that have higher order relationships between them.
3. Although some variables might be the most informative in a particular dataset, adding new information could make those variables less relevant.

Problem 4: Air Traffic Time Series Model

1 Introduction

The purpose of Part 4 of this report is: (1) to explain the methodology used to develop three models used for the prediction of air traffic (boardings, passenger miles, and freight ton miles) during the months of September, October, November, and December of 2014; (2) to analyze the performance of such models when training and testing on the dataset provided; and (3) to present the predictions for the abovementioned months. The following questions will be answered:

1. What can be said about the friend's work and advice?
2. What engineered features were created and why each of one?
3. Which engineering features created seem to aim in the prediction?
4. What cross validation process was used and why that one?
5. What is the expected performance of the models?
6. What are the values for the final predictions on the months of interest?

2 Dataset

The dataset provided is a time series from January 1996 to August 2014 with the monthly passenger *boardings, revenue passenger miles,* and *freight ton miles* for all flights in the US. Additionally, for the same period the dataset has the *average monthly temperatures, crude oil spot price,* and *GDP on a quarterly basis.* A sample of the time series is shown in Table 4.1 (note that year values start at 1 for 1996, 2 for 1997 and so on). Information on the average temperature, crude oil spot price and GDP is also available (at least as a prediction) for the months to be predicted.

Table 4.1: Sample data from "Permits 2014"

year	month	oil	temp	gdp	boardings	revenue	freight
1	Jan	29.59	31.35	10508.1	49618	52049391	1394321
1	Feb	29.61	33.98	10508.1	48129	48845243	1442846
1	Mar	27.25	41.49	10508.1	59121	60769074	1661091

3 Analysis

The analysis will be divided in two parts: the first section will cover the analysis on the advice and method provided by the "friend"; the second section will explain the methodology for manipulating the data and creating the features to be used as explanatory variables; and the last part will present the results from the models, as well as the final predictions.

3.1 EDA and Comments on Friend's Work and Advice

Considering that the dataset consists of a time series, the first step towards exploring it was to plot values against time. Figure 4.1 shows *boardings, passenger miles traveled (*from now on referred to as *revenue*), and *tons of freight* (from now on called just *freight*), as a function of months. It can immediately be observed that there are dependencies on all three variables depending on which month you are looking

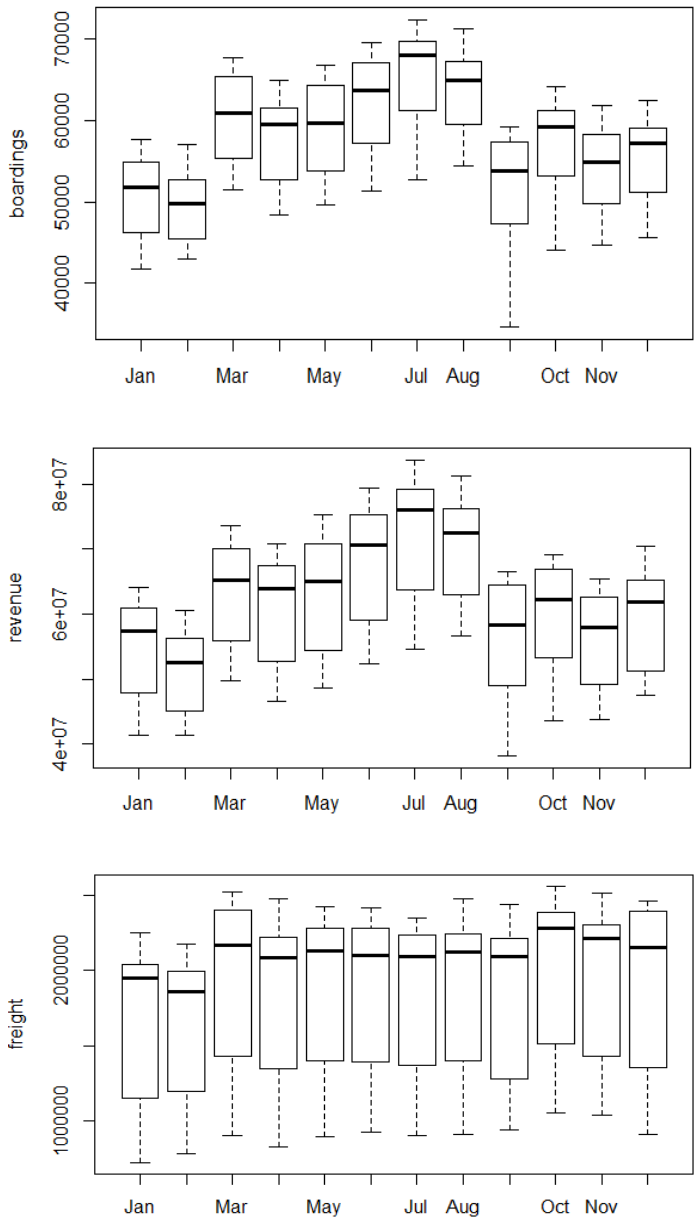

Figure 4.1: Boardings, Revenue and Freight as a function of months

at. These dependencies are more pronounced in boardings and revenue, and although freight values tend to be more constant throughout, they also have a larger variance within each month.

The behavior of all three variables is also very dependent on time across the set of years in the dataset. In Figure 4.2 an increasing tendency can be deduced on the three variables, and it can also be seen how similar boardings and revenue are. This is no coincidence, as revenue is actually a function of boardings (boardings times the number of miles traveled). The behavior of freight is a little more unusual and is more susceptible to macroeconomic factors that are not included in the dataset (through some online research I found that the jump observed at about 80 on the x-axis corresponds to the opening of new air channels between the US and Asian countries like China and Singapore).

To finalize the data exploration, a pairs plot was done to observe the relationship among the variables. As pointed out earlier, there is a very strong positive correlation between boardings and revenue. Another highly correlated pair of variables are oil prices and GDP. Slightly weaker, but still significant correlations can be observed between GDP and the three variables to be predicted. Oil has a similar structure when plotted against the predictable variables. Finally, temperature exhibits a similar trend, but with a much higher variability.

The same methodology used by Bob (that is my friend's name) was used as a first step to solve Problem 4. As mentioned above, there are strong correlations between the explanatory and predicted variables. A multiple linear regression model was also developed and although the results were not as favorable as those suggested by Bob—mainly because I did not perform any feature engineering at this point—it is clear that a simple model can be developed in this fashion just by looking at how correlated some of the outputs are. There are, however, two main problems that Bob's model could have:

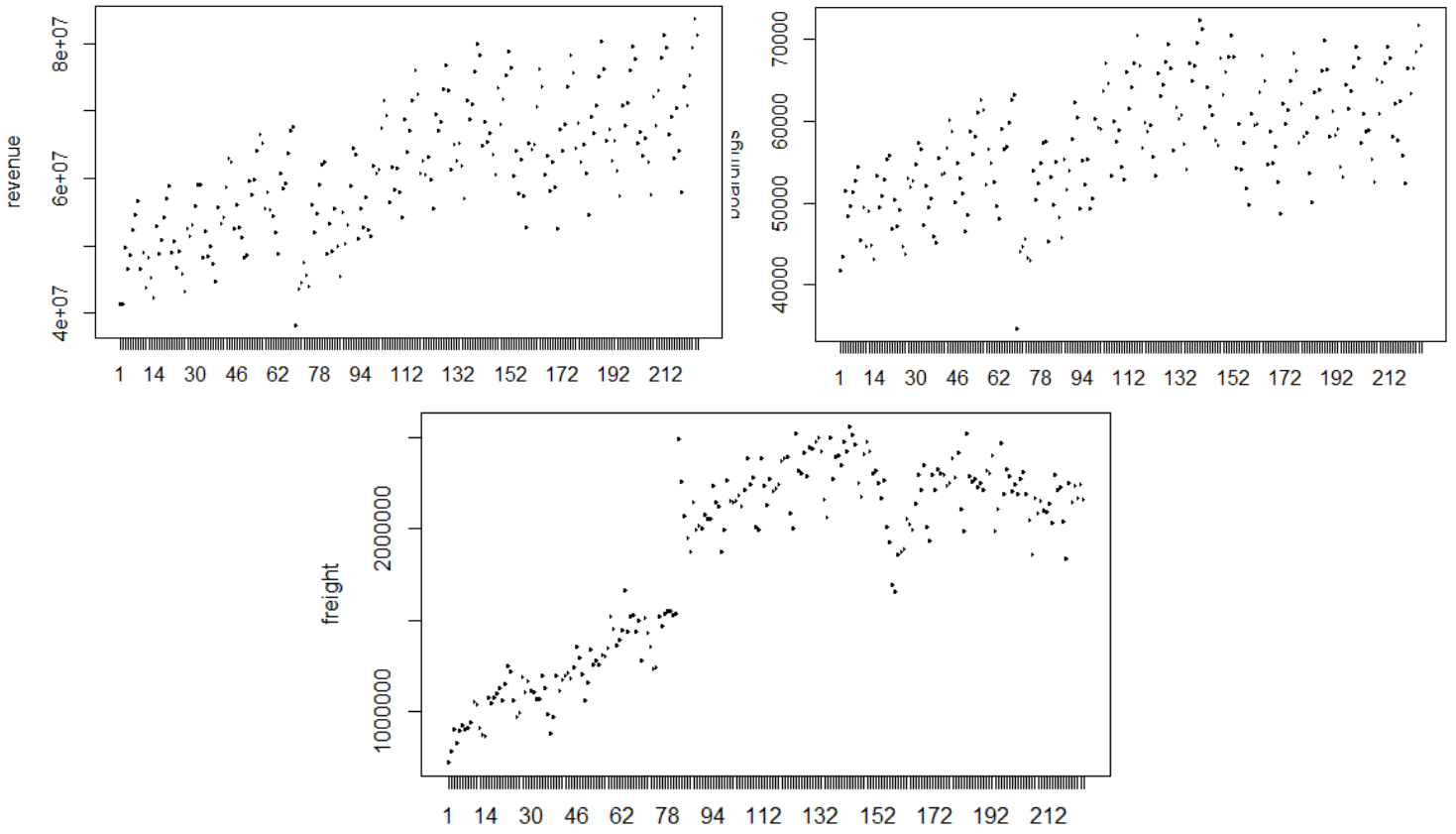

Figure 4.2: Boardings, Revenue and Freight as a function of time (numbers on x-axis represent a month

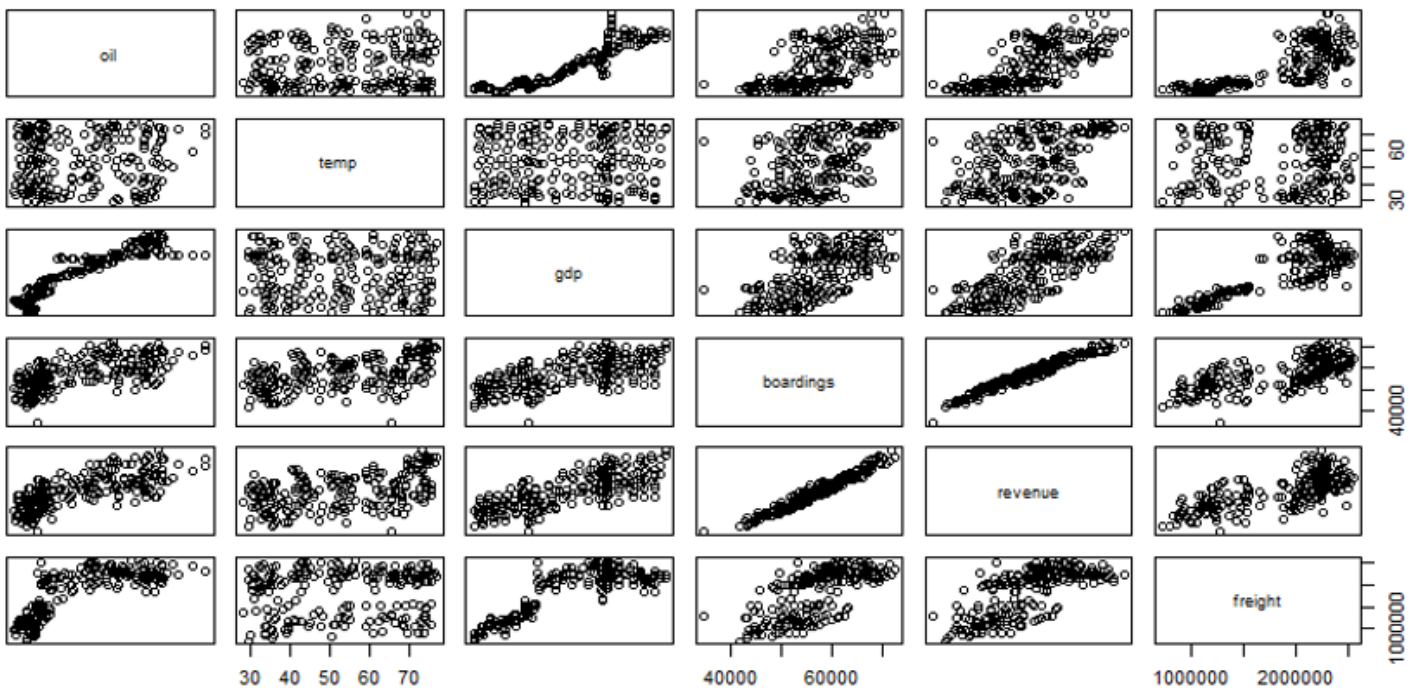

Figure 4.3: Pairs plot of (going from left to right): oil prices, temperature, GDP, boardings, revenue, and freight

1. The model described by Bob may provide low p-values, high adjusted R-squared values, and produce an excellent Residuals plot, however, it may be overfitting the dataset. The main issue with Bob's model is that although it performs great when confirming the data, it lacks true predicting power because we don't really know how well it does on data that it has not seen before. Meaning, it has not been tested on data that was not used to fit the model. Moreover, the dataset does not include any information on those months from Sept to Dec 2014, only predictions. That means that the error could be even greater. Finally, the error in Bob's model would not take into account the fact that we are predicting in future values, as the time relationships, or the error resulting from them, could be missed by the model.
2. A linear model may not be accounting for underlying structures in the data, which are not evident when fitting the dataset provided, but may become more apparent when trying to predict on previously unseen data.

After analyzing the linear regression model, and trying different models of different complexities, I decided to use a Multivariate Adaptive Regression Spline (MARS) for my final models. The methodology and results are provided next.

3.2 Data Manipulation, Features Engineering, Model Selection, and Cross Validation Process

The data was first manipulated and grouped into one single data frame. Having done that, some feature arrangements and engineering was done. First, the direct factors were chosen, which ended up being:

1. Boardings, Revenue and Freight values of the same month in the previous year.
2. Boardings, Revenue and Freight values of the previous quarter (current month minus four months)
3. Temperature of last quarter (ideally it would have been of last month but Oct through Dec of 2014 would not have had a value for it).

Then, the engineered features were created in a combination of different forms. Key things that I considered were the fact that values have a highly dependent on which month you are looking at, and also that some variables—such as oil prices and GDP—have high correlations that could make it redundant to use both, for instance. The most useful engineered features tried were the following:

1. **avg_boardings -** Boardings, Revenue and Freight average values over the previous 4 years prior to the year before.
2. **prev_month_temp -** Average between the temperatures of months -4 and -5 from current month.
3. **prev_month_gdp.oil -** GDP to Oil Price Ratio for the previous quarter.
4. **last_year_miles -** Miles traveled last year during the same month (obtained by dividing revenue over boardings).
5. **last2_ratio_boardings (or revenue or freight) -** Boardings, Revenue and Freight values of previous year times the ratio obtained by dividing Boardings, Revenue and Freight values from the previous year over those from two years before. This was done to account for the increase in these values that was observed from the time series plots. It is like multiplying the previous year by a coefficient that is slightly greater than one and accounts for the most recent increase in air traffic on a per year basis.

6. **boardings_gdp (or revenue or freight) -** Boardings, Revenue and Freight values for the same month on last year, divided by the GDP of the same month from last year, and multiplied by the GDP of the previous quarter. This was done to have a ratio on the increment in GDP that was experienced over the last year, and multiply that by the volumes of air traffic observed during the same months on the previous year. Therefore, air traffic would be adjusted to some extent to a more recent economic development or downturn, which was shown via EDA to have a strong relationship with air traffic volumes.

Having obtained the features, they were all compiled into a large dataset. This dataset was used for training, testing, and later on to train again on the entire dataset and predict on the last few months of interest. The only missing values from this whole procedure were those from year 1996 to 2000, because of those features that were obtained by averaging values over the previous five years.

The fitting and testing split was a very simple one, which was randomized over the entire generated dataset described above. An alternative way of fitting and testing the data would have been to train on previous years and test on later years, so that the error in the prediction would already account for the increments of values in all three variables as time progresses. This could provide us with a better idea on how good the model is in predicting future values in time. However, the simpler method was chosen as time was a key component that I accounted for when generating the engineered features, and therefore I hope these time relationships are already accounted for to a considerable extent.

At first, the desired cross validation was one where the model would fit on earlier years and validate on future ones, to account for the value increments over time. Also, having a 12-fold (one for each month) could have been very useful on training the model to account for differences between months. However, as the cross validation was performed automatically by the program, there were a great deal of limitations as to how the whole process would be carried on. I tried different splits, from 6 to 10-fold, and different repetition values, from 2 to 5. In the end the results were very similar between different forms, and the one I ended up using was the more typical one of 10-fold and doing 5 repetitions. The obtained results are provided next.

3.2 Results and Performances

The methodology for predicting Boardings, Revenue and Freight was to create one model for each variable. Each model consisted of different explanatory variables, as well as different parameters chosen as the optimal ones by the algorithm of the MARS model. The value of these parameters are shown in Table 4.2.

Also shown in Figure 4.2 are the RMSE and the RMSE to Mean ratio, which was obtained by divining the RMSE over the average value across the entire dataset of each of the dependent factors. The RMSE was obtained by predicting on the test data after fitting on the training part of the data. All three error-to-mean ratios are considerably low, and they were obtained after performing a forward variable selection process.

Table 4.2: Characteristics and performance for each of the three models (boardings, revenue and freight)

	Nprune	Degree	RMSE	RMSE/Mean
Boardings	11	1	2,139	3.60%
Revenue	12	1	3,168,166	4.93%
Freight	13	1	118,503	5.61%

During the variable selection process, it was observed that including some of the indirect (engineered) features provided enhanced performances in the models. Those models that provided, respectively, the best results, based on the lowest RMSE, were:

1. **boardings = boardings_gdp + gdp + month**
2. **revenue = month + temp + gdp + last_year_revenue**
3. **freight = month + oil + temp + gdp + last_year_freight + avg_prev_temp + freight_gdp**

As it can be observed, all of them included a combination of direct and indirect features, signaling the benefits from using EDA understand the relationships and patterns in the data, to then perform feature engineering. Once the models were fitted and tested, I proceeded to train them on the entire dataset. After doing so, they were used to predict on the missing values from September 2014 to December 2014. The final obtained predictions are presented in Table 4.3 below.

Table 4.3: Model predictions on Boardings, Revenue and Freight for the months of Sept – Dec of 2014

	Boardings	Revenue	Freight
Sep-14	58,656	67,563,004	2,092,554
Oct-14	61,998	68,056,959	2,228,945
Nov-14	58,764	65,198,539	2,103,213
Dec-14	62,042	69,336,154	2,056,409

4 Conclusions

Although a simple linear model was developed originally to fit the data, and the results were very acceptable, the nature of predicting models requires for analysts to train on certain data to then test on data that has not been "seen" by the model. This is done mainly to prevent the development of models that overfit the available data but perform poorly in new data. This is also done to know what the error of the model is when predicting in "real" data, which allows us to know how good the model really is, and take the results for what they are: approximations with an inherent error. Ultimately, instead of having a model that outputs a number, we are thus able to have a model that provides estimates where we know what the confidence intervals are and whether or not the results are statistically significant.

Another important feature of predictive data analytics, especially when dealing with time series, is that feature engineering is a crucial component on the development of a model. A dependent variable can be expressed as a function of many factors, even more so when time is one of them, but this does not mean all of them will be informative. Moreover, through different forms of cross validation, one can create models that are trained to account for the uncertainty and variability that comes from having *time* as a factor (constantly increasing or decreasing variables, periodicity, lags, etc.). By manipulating the given factors I attempted to account for those underlying structures and patterns, and I developed three models that performed better than any combination of the direct features. This is not to say, however, that better models could not be created. As a matter of fact, trying more combination of factors, or bringing external ones into the equation (such as the expertise from someone knowledgeable in the field) would definitely be beneficial in finding superior models.