**TESLA PROJECT PAPER**

**1. Project Proposal and Presentation**

• What is the problem you are trying to solve or question you are trying to answer?

• What data do you need?

**• What work do you plan to do in the project?**

In our project we will focus on analyzing Tesla sales and how they are impacted by external factors namely the fluctuations in oil prices and the money spent on other electric cars in general.

**• Which algorithms/techniques/models do you plan to use/develop? Be as specific as you can?**

We will do so by running a multi linear regression model using the monthly sales of Tesla as the response variable and the monthly oil prices and money spend on substitute electric cars as explanatory variables.

**• How will you evaluate what you’ve done?**

We will first analyze and observe our data to see whether it meets the condition to follow all the assumptions of a linear regression model such as population size, normality and minimized skewness.

**• What do you expect to submit/accomplish by the end of the project?**

We expect to define a clear relationship between the sales fluctuation of Tesla cars and the movements of the external factors mentioned above. Ideally our model will be strong and sound and provide insights into what drives the sales performance of Tesla.

The project proposal should follow the guidelines provided by the IEEE at: <http://www.ieee.org/publications_standards/publications/authors/authors_journals.html>

**2. Status Report**

• What the problem is that you are trying to solve or question you are trying to answer.

• All relevant background information including any relevant literature you have/will use.

• The overall process you will follow for the entire project.

• A description of your data including how you obtained it.

**• A description of any relevant, interesting exploratory data analyses.**

We looked into the Tesla stock prices and how they fluctuate given oil prices. We considered using the tesla stock price as a proxy for the tesla sales performance but after reviewing the data we decided to use the actual tesla sales. Indeed, the movements in tesla stock price are affected by factor beyond our analysis such as the market in general and the sentiment of the general population about the company and the brand.

**• A description of the methods/techniques/tools/algorithms you have/will use to complete the project.**

We decided to focus on tesla monthly sales from June 2012 to March 2016 excluding July 2013 because we could not find the data for this time period. This gave us a data set of 45 observations.

**• A description of the challenges you have had working on the project.**

The biggest challenge was finding the data needed in a format that was logical and made sense given the model we needed to run. We had sales data, oil prices and substitutes sales data but all for different periods and in different breakdown. We had daily prices and yearly sales and monthly data elsewhere as well. We decided to normalize every data to have everything on a monthly basis so that our dataset could contain enough observations to ensure that any model we come up with was sound.

**• A discussion of the parts of the project that have been completed.**

We completed our exploratory data analysis and chose the variables needed for our model.

**• A discussion of the parts of the project that remain to be completed.**

We still need to finalize the format of our dataset in order to run the model and discuss our findings.

**• A discussion of how you will finish the final project report and presentation.**

???

The status report should follow the guidelines provided by the IEEE at: <http://www.ieee.org/publications_standards/publications/authors/authors_journals.html>

**3. Final Project Report**

• Introduction, motivation and general description of the situation, problem or challenge. o Following the proposal and status report, what is the situation, problem or challenge you are addressing?

o What preliminary examination leads you to believe analytics could help?

o What are the shortcomings of the current work/analysis that analytics could help with?

• Related work. o Provide a thorough background for the project; e.g. about the company, about the situation, problem or challenge, about other companies that have undergone similar situations, problems or challenges and how they handled them or did not, etc.

o How does this project relate to other work that has been done on this situation, problem or challenge?

Provide the appropriate citations/references per the author instructions at: <http://www.ieee.org/publications_standards/publications/authors/authors_journals.html>

• Data o Give a complete description of the data you use during the project, including any you reject. Provide the source(s) of your data.

Provide a detailed description of your data.

Provide any exploratory data analyses you complete.

**• Technical Approach**

**o Give a detailed description of the process for your entire project.**

After collecting and reorganizing the data we have a final dataset with 45 observations for the monthly sales of Tesla, the monthly oil prices, and the substitute unit prices for the same period (June 2012 to March 2016). In order to remove the impact of the difference in scale we standardize the data and run our model with the standardized dataset.

Then we run a regression using the tool R and the standardized oil prices and substitute prices as explanatory variables and the standardized tesla sales as response variable. Here is the resulting model from our first iteration:

T.Sales = -1.055e-10 – 0.14 x oil price + 0.48 x substitute.unit.price

For this first iteration, an evaluation of the model using the F-test gave us:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.055e-10 1.226e-01 0.000 1.0000

oil.price -1.383e-01 2.798e-01 -0.494 0.6238

Substitute.unit.price 4.796e-01 2.798e-01 1.714 0.0939 .

---

Residual standard error: 0.8227 on 42 degrees of freedom

Multiple R-squared: 0.3683, Adjusted R-squared: 0.3382

F-statistic: 12.24 on 2 and 42 DF, p-value: 6.474e-05

So even if the model as a whole is significant the oil price variable is not significant to the model so we rerun it without the oil price and here is the result:

T.Sales = -9.812e-11 + 0.6 x substitute.unit.price

This second iteration is significant as a whole and has the substitute unit price significant as well – with a much lower p-value:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -9.812e-11 1.216e-01 0.000 1

Substitute.unit.price 6.038e-01 1.216e-01 4.967 1.13e-05 \*\*\*

---

Residual standard error: 0.8154 on 43 degrees of freedom

Multiple R-squared: 0.3646, Adjusted R-squared: 0.3498

F-statistic: 24.67 on 1 and 43 DF, p-value: 1.127e-05

We even notice a slight increase of the Adjusted R-squared from 34% to 35%

From this analysis we can say that given our dataset, each unit increase in the standardized substitute prices give us an increase of 0.6 units in the standardized tesla monthly sales. The intercept being extremely minimal, this means that when the substitute sales are zero, according to this model, the tesla sales are zero as well. Keep in mind that the intercept is not significant (p-value is 1) but this is often the case in regression analyses.

We got a value of 35% for the Adjusted R-squared. This means that 35% of the variability in the tesla monthly sales is explained by the model we obtained. Therefore, it is safe to assume that there are still several other factors outside of this analysis that are responsible for the movements of tesla monthly sales. Given the fact that our value of the Adjusted R-squared is quite low, we can’t attribute all the remaining factors to random events.

After obtaining a model that shows us the relationship between tesla sales and substitute unit prices we can check whether this relationship is significant. Our hypothesis is that there is a significant positive relationship between tesla sales and the unit prices of substitute cars.

The F-test results already confirmed to us that the relationship is significant (p-value is minimal). Then, we compute a 95% confidence interval for the coefficient of the substitute unit price.

The result is: [0.35 ; 0.85]. Therefore, we are 95% confident that one unit increase in the standardized substitute unit price leads to an increase ranging from 0.35 to 0.85 units in standardized tesla sales. Since the 95% confidence interval does not include zero it is safe to say that there is a positive relationship between tesla sales and substitute unit price. This makes sense and follow our logic since an increase in the price of a substitute product would most likely lead to an increase in sales for that product.

**• Test and evaluation**

**o Describe how you test your approach to ensure that it is valid.**

We test our approach and ensure that it is valid by observing the distribution of the tesla sales and see whether or not it is normal or nearly normal. Our dataset is the “entire population” since it includes every sales (that we could reasonably account for) since the creation of Tesla. This dataset counts 45 observations, which by the rule of thumb is considered to be a large-sized dataset since the size is greater than 30.

**o Discuss the validity of your approach.**

Now looking at the histogram and the box plot we see that the sales distribution is right skewed with three outliers. We could have removed the outliers in order to perform the analysis with a dataset with a better fit to a normal distribution but we chose not to. These 3 outliers actually represent the most recent sales numbers. Despite the fact that they are outliers in this particular dataset they represent the beginning of an upward trend for Tesla sales. We expect these values to become more mainstream in the coming months so we chose to keep them in the model to run the analysis.

Given the fact that the size of our dataset is large enough and the distribution shows some skewness but not extensive, we proceed with our analysis by considering our dataset to be nearly normal.

**o Describe how you will evaluate your results and/or conclusions including any specific metrics, output data, completed analyses, etc.**

already discussed earlier right(?)

**o Discuss the baseline you will use to compare your results to.**

already discussed earlier right(?)

**o Discuss how well your approach worked to address the situation or challenge, solve the problem or answer the research question.**

As a summary we can say that our analysis:

* Tesla sales nearly normal with very recent upward trend
* Was good start but not complete for lack of other variables that could improve the model
* Support our hypothesis as it show shows significant positive relationship between substitute unit price and tesla sales

**o Discuss any potential future work. For example, if you were not able to resolve the situation or problem or answer the research question what will it take to do so? What else needs to be done?**

Our model was sound in itself but the adjusted R-squared obtained tells us that there is more to the story. So any potential future work would look into which variables would add more value to the model that can better describe the sales performance of Tesla. One potential example might be the sales of cars that Tesla considers to be its competition. Indeed, while our analysis was conducted thinking of Tesla cars to be part of the greater group of electric cars, the company associates itself more with luxury than with electric / hybrid vehicle manufacturers. Therefore an analysis that would include sales information on such competitors could bring more value to the model.

o Evaluate and report whether or not someone unfamiliar with your work could accurately replicate it.

¯\\_(ツ)\_/¯ lol

The final report should follow the guidelines provided by the IEEE at: http://www.ieee.org/publications\_standards/publications/authors/authors\_journals.html.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

CODE

> TESLA.DATASET <- read.csv("C:/Users/clementine.djouka/Desktop/HARRISBURG/SPRING 2016/ANLY 500/PROJECT/TESLA DATASET.csv")

> View(TESLA.DATASET)

> lm(T.sales ~ oil.price + substitute.unit.price, data = TESLA.DATASET)

Error in eval(expr, envir, enclos) :

object 'substitute.unit.price' not found

> lm(T.sales ~ oil.price + Substitute.unit.price, data = TESLA.DATASET)

Call:

lm(formula = T.sales ~ oil.price + Substitute.unit.price, data = TESLA.DATASET)

Coefficients:

(Intercept) oil.price Substitute.unit.price

-1.055e-10 -1.383e-01 4.796e-01

> mlr(T.sales ~ oil.price + Substitute.unit.price, data = TESLA.DATASET)

Error: could not find function "mlr"

> model <- lm (T.sales ~ oil.price + Substitue.unit.price, data = TESLA.DATASET)

Error in eval(expr, envir, enclos) :

object 'Substitue.unit.price' not found

> model <- lm (T.sales ~ oil.price + Substitute.unit.price, data = TESLA.DATASET)

> summary(model)

Call:

lm(formula = T.sales ~ oil.price + Substitute.unit.price, data = TESLA.DATASET)

Residuals:

Min 1Q Median 3Q Max

-1.42412 -0.61400 0.06834 0.40527 2.99164

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.055e-10 1.226e-01 0.000 1.0000

oil.price -1.383e-01 2.798e-01 -0.494 0.6238

Substitute.unit.price 4.796e-01 2.798e-01 1.714 0.0939 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8227 on 42 degrees of freedom

Multiple R-squared: 0.3683, Adjusted R-squared: 0.3382

F-statistic: 12.24 on 2 and 42 DF, p-value: 6.474e-05

> lm(T.sales ~ Substitue.unit.price, ddata = TESLA.DATASET)

Error in eval(expr, envir, enclos) : object 'T.sales' not found

> lm(T.sales ~ Substitute.unit.price, data = TESLA.DATASET)

Call:

lm(formula = T.sales ~ Substitute.unit.price, data = TESLA.DATASET)

Coefficients:

(Intercept) Substitute.unit.price

-9.812e-11 6.038e-01

> model2 <- lm(T.sales ~ Substitute.unit.price, ddata = TESLA.DATASET)

Error in eval(expr, envir, enclos) : object 'T.sales' not found

> model2 <- lm(T.sales ~ Substitute.unit.price, ddata = TESLA.DATASET)

Error in eval(expr, envir, enclos) : object 'T.sales' not found

> model2 <- lm(T.sales ~ Substitute.unit.price, data = TESLA.DATASET)

> summary(model2)

Call:

lm(formula = T.sales ~ Substitute.unit.price, data = TESLA.DATASET)

Residuals:

Min 1Q Median 3Q Max

-1.35239 -0.65071 0.06305 0.42856 3.06545

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -9.812e-11 1.216e-01 0.000 1

Substitute.unit.price 6.038e-01 1.216e-01 4.967 1.13e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8154 on 43 degrees of freedom

Multiple R-squared: 0.3646, Adjusted R-squared: 0.3498

F-statistic: 24.67 on 1 and 43 DF, p-value: 1.127e-05

> hist(TESLA.DATASET$T.sales)

> boxplot(TESLA.DATASET$T.sales)

> f.test(model2)

Error: could not find function "f.test"

> abline(model2)

> inference(y = TESLA.DATASET$T.sales, x = TESLA.DATASET$Substitute.unit.price, est = "mean", type = "ci", null = 0,

+ alternative = "twosided", method = "theoretical")

Response variable: numerical, Explanatory variable: categorical

Error: Categorical variable has more than 2 levels, confidence interval is undefined, use ANOVA to test for a difference between means.

In addition: Warning messages:

1: package ‘BHH2’ was built under R version 3.2.4

2: Explanatory variable was numerical, it has been converted to categorical. In order to avoid this warning, first convert your explanatory variable to a categorical variable using the as.factor() function.

> var.test(model2)

Error in var.test.default(model2) :

argument "y" is missing, with no default

> var.test(T.sales ~ Substitute.unit.price, data = TESLA.DATASET)

Error in var.test.formula(T.sales ~ Substitute.unit.price, data = TESLA.DATASET) :

grouping factor must have exactly 2 levels

> var.test(T.sales ~ Substitute.unit.price)

Error in eval(expr, envir, enclos) : object 'T.sales' not found

> var.test(TESLA.DATASET$T.sales , TESLA.DATASET$Substitute.unit.price)

F test to compare two variances

data: TESLA.DATASET$T.sales and TESLA.DATASET$Substitute.unit.price

F = 1, num df = 44, denom df = 44, p-value = 1

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.549539 1.819707

sample estimates:

ratio of variances

1

> confint(model2, Substitute.unit.price)

Error in confint.lm(model2, Substitute.unit.price) :

object 'Substitute.unit.price' not found

> confint(model2, "Substitute.unit.price")

2.5 % 97.5 %

Substitute.unit.price 0.3586706 0.8489682



