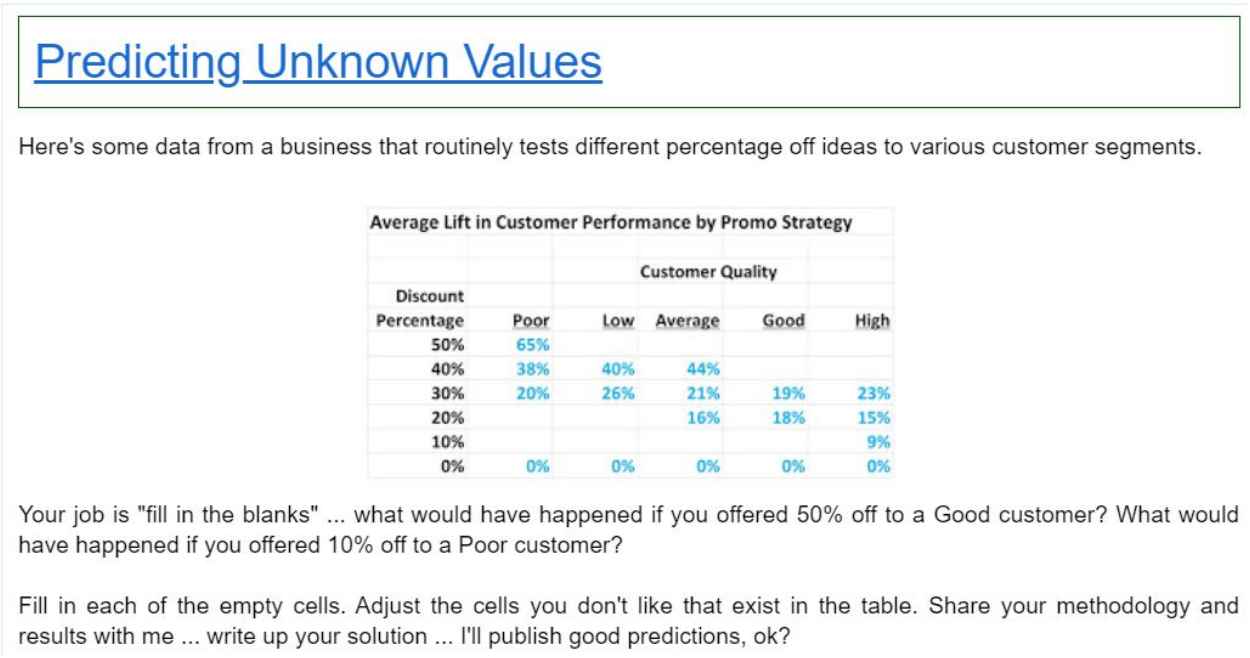
Predicting Unknown Values Quiz

July 2020

EXECUTIVE SUMMARY

BACKGROUND

The team was asked to solve the quiz on Figure 1:



Predicting Unknown Values

Here's some data from a business that routinely tests different percentage off ideas to various customer segments.

Average Lift in Customer Performance by Promo Strategy

	Customer Quality				
Discount Percentage	Poor	Low	Average	Good	High
50%	65%				
40%	38%	40%	44%		
30%	20%	26%	21%	19%	23%
20%			16%	18%	15%
10%					9%
0%	0%	0%	0%	0%	0%

Your job is "fill in the blanks" ... what would have happened if you offered 50% off to a Good customer? What would have happened if you offered 10% off to a Poor customer?

Fill in each of the empty cells. Adjust the cells you don't like that exist in the table. Share your methodology and results with me ... write up your solution ... I'll publish good predictions, ok?

Figure 1: The quiz.

The quiz covers 5 groups of customers separated by their quality: Poor, Low, Average, Good, and

High. Various discount rates are offered to all groups: 0%, 10%, 20%, 30%, 40%, 50%.

The quiz departs from the base scenario where no discount is offered to the customers. Further, the customers are seduced by discounts and the table reflects the customers' reaction to the stimulus. For example, the High customers' performance improved by 9% when the discount reached 10%.

The team decided to stick to the original numbers and undertook no adjustments.

PREDICTED VALUES

Employing the subsequently described workflow, the team reached the following set of predictions highlighted in gray color.

	poor	low	average	good	high
50%	65%	69.4%	69.0%	67.6%	67.9%
40%	38%	40%	44%	38.5%	38.8%
30%	20%	26%	21%	19%	23%
20%	13%	16.4%	16%	18%	15%
10%	7.5%	10.6%	10.2%	8.8%	9%
0%	0%	0%	0%	0%	0%

Typical CRISP/ASUM-DM protocol was applied to reach above mentioned predictions. Workflow employed consists of business & data understanding, data quality and sanity checks, exploratory data analysis (statistical tests and visualizations), model fitting, model performance assessment, model selection and finally predictions.

WORKFLOW

A. Dataset Preparation & Presentation

- The original data were transformed into a data frame suitable for further on analysis/modeling.
- Moreover the dataset is split into train and test based on outcome availability. We are aware of the challenge to predict 12 instances on the basis of an 18 instance training sample.
- Discount: A numerical variable, with observed values in the range [0 - 0.5] denoting the amount of discount of a promo to a segment of the customers
- Customer Quality: a categorical variable, with five (5) unique values [poor, low, average, good, high] denoting customers segmentation.
- Average Lift: a numerical variable, with observed values in the range [0 - 0.65], denoting the on average increase in purchasing activity a group of customers presented as a result of a specific discount promo.

1. Train Dataset

The train sample is presented in Table 1.

	discount	cust_quality	avg_lift
1	0.5	poor	0.65
2	0.4	poor	0.38
3	0.3	poor	0.20
6	0.0	poor	0.00
8	0.4	low	0.40
9	0.3	low	0.26
12	0.0	low	0.00
14	0.4	average	0.44
15	0.3	average	0.21
16	0.2	average	0.16
18	0.0	average	0.00
21	0.3	good	0.19
22	0.2	good	0.18

24	0.0	good	0.00
27	0.3	high	0.23
28	0.2	high	0.15
29	0.1	high	0.09
30	0.0	high	0.00

Table 1: Training sample.

2. Test Dataset

The Test Dataset is presented in Table 2.

	discount	cust_quality	avg_lift
4	0.2	poor	NA
5	0.1	poor	NA
7	0.5	low	NA
10	0.2	low	NA
11	0.1	low	NA
13	0.5	average	NA
17	0.1	average	NA
19	0.5	good	NA
20	0.4	good	NA
23	0.1	good	NA
25	0.5	high	NA
26	0.4	high	NA

Table 2: Test Dataset.

B. Data Quality and Sanity Checks

No need to do this with code cause based on simple visual inspection, we see that:

- No missing values on the input data (missing values in the outcome reflect the target to compute)
- No duplicated records
- No irregular values (discounts, customer quality types as well as outcomes are within reasonable math/business values spectrum)

C. Exploratory Data Analysis

Insights:

- We observe that the higher the discount, the higher the average lift what makes sense to us. We expect the prediction to preserve this pattern.
- Maybe a non linear model could explain a bit better the observed relationships in discount ~ avg_lift, however precautions against overfitting should be applied what measures we applied to tackle a potential overfitting.
- The higher the customer quality the lower the avg_lift impact due to some discount, giving the impression that higher quality customers are less affected by discounts and we expect the prediction to follow this pattern
- Attention: could be also attributed to the available data (high quality customers are missing high discount avg_lift, therefore results here could be spurious)
  Maybe replace below the “correlation plot” with “scatter plot”

Any

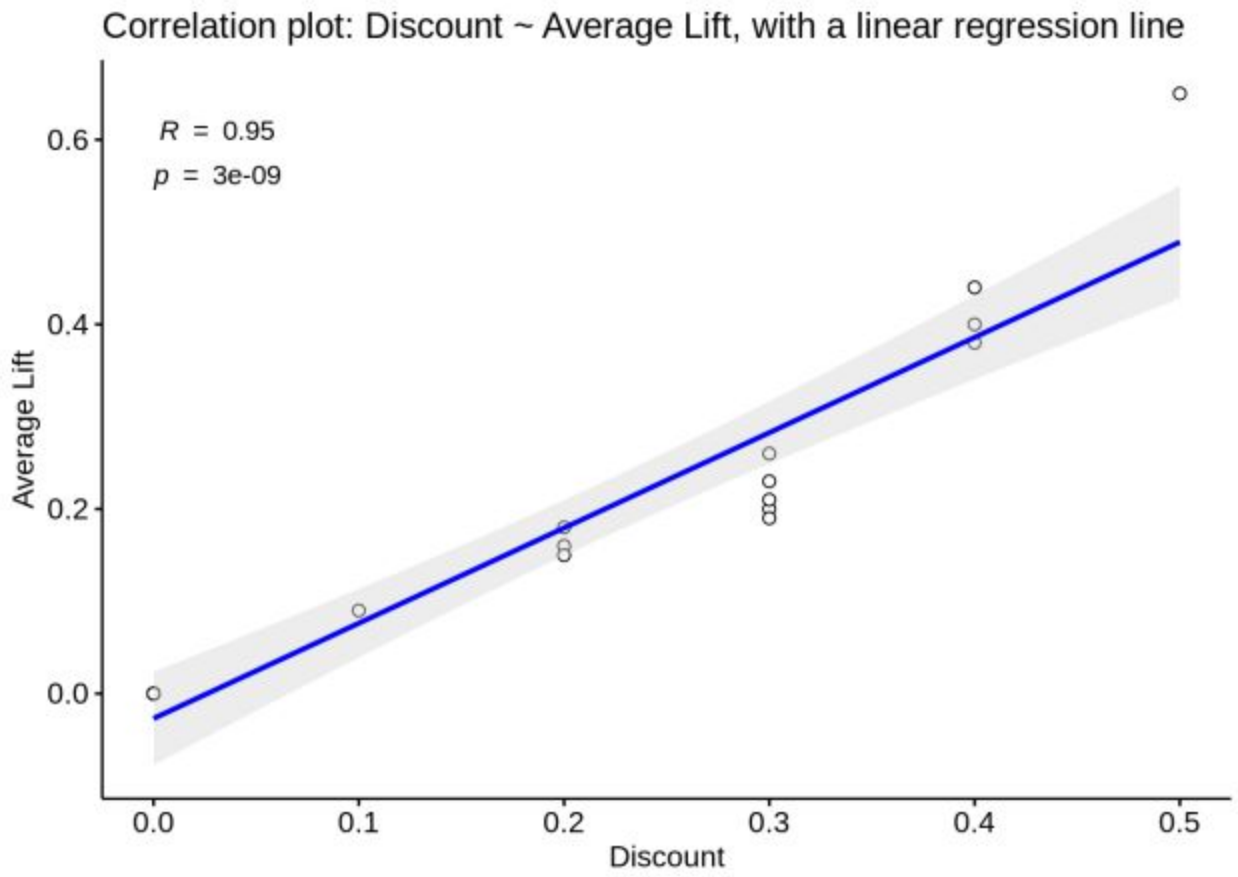


Figure 2: Discount ~ Average Lift linear relationship.

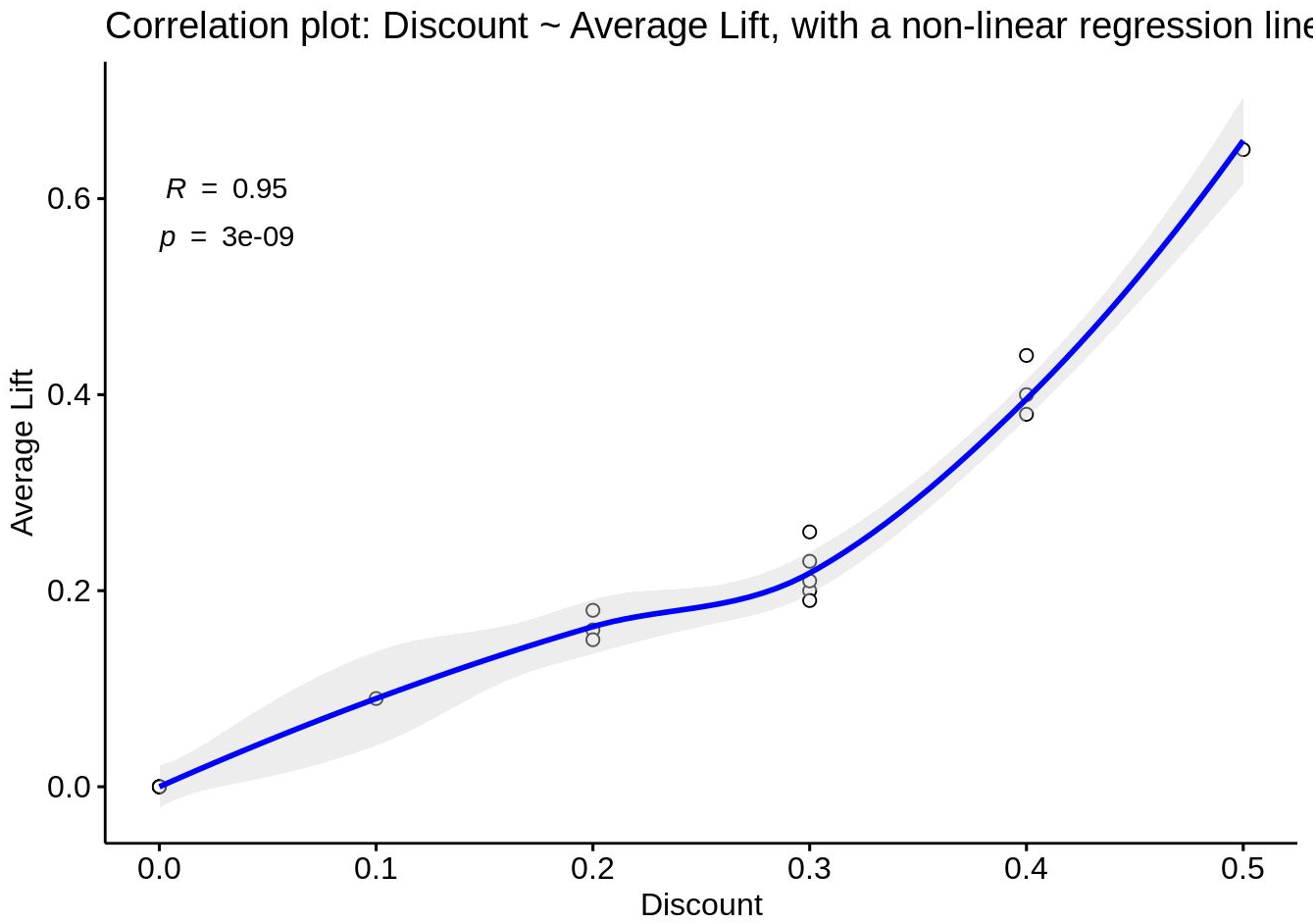


Figure 3: Discount ~ Average Lift non-linear relationship.

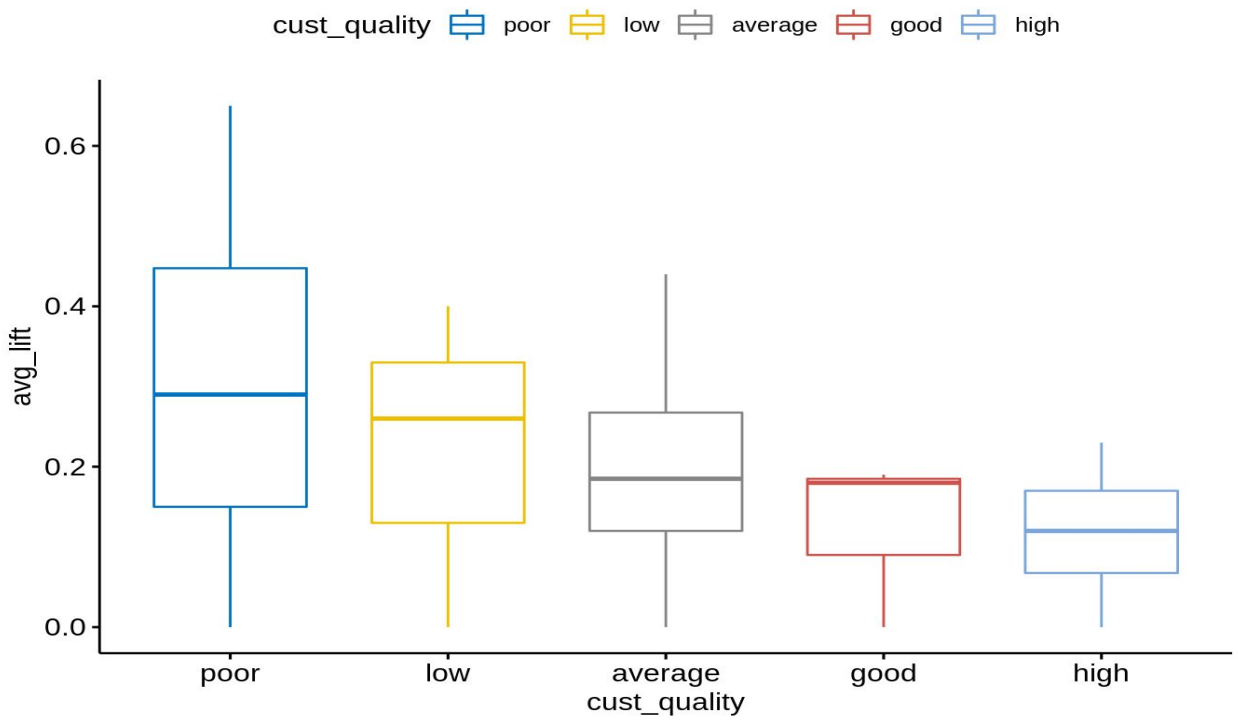


Figure 4a: Customer Quality ~ Average Lift boxplots

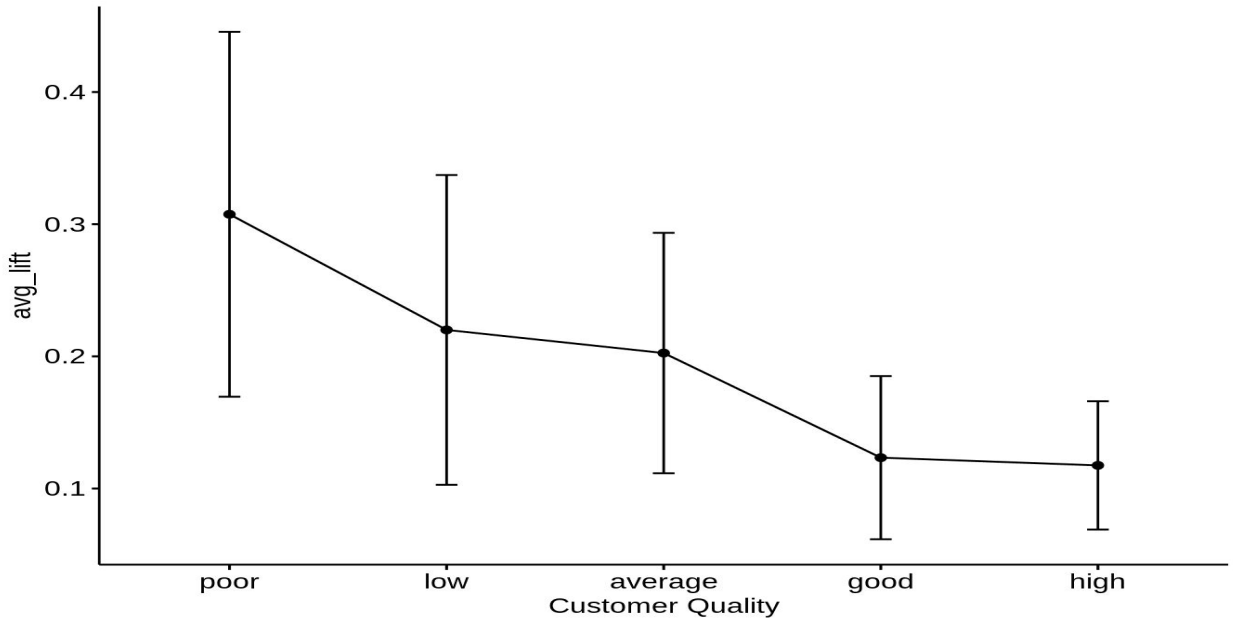


Figure 4b: Customer Quality ~ Average Lift mean + variance per group

D. Modeling & Model Selection

- The team decided to apply linear and non-linear regression, and decision tree methods, in order to cross-check the method's performance and on this basis select the best method. To tackle the limited training sample (18 dots only), we decide to resample. The corresponding models are fitted to the available training set under a resampling scheme (9 fold cross validation (due to 18 train data points)) so as to get an estimate of performance along with the respective uncertainty estimate.
- Among the available fit measures the team selected the MAE. Table 3 reads the performance results based on a 9 fold cross validation scheme.
- Besides the visualizations on Figure 6 and Figure 7, a hypothesis test for the two best performing models is run. Occam's razor rule shall be applied in case the hypothesis test does not appoint a clear winning model.

```
##
## Call:
## summary.resamples(object = rValues)
##
## Models: LinearRegression, nonLinearRegression, DecisionTree
## Number of resamples: 9
##
## MAE
##                          Min.    1st Qu.     Median       Mean    3rd Qu.
## LinearRegression    0.034785012 0.03762739 0.05666667 0.06430388 0.08568182
## nonLinearRegression 0.001152959 0.01688891 0.03599982 0.03146454 0.03813218
## DecisionTree        0.010000000 0.05000000 0.13687500 0.14256944 0.19000000
##                          Max. NA's
## LinearRegression    0.12065882    0
## nonLinearRegression 0.07673077    0
## DecisionTree        0.39187500    0
##
## RMSE
##                          Min.    1st Qu.     Median       Mean    3rd Qu.
## LinearRegression    0.037842443 0.04035270 0.06713502 0.07411515 0.08944792
## nonLinearRegression 0.001161241 0.01699483 0.03901991 0.03596031 0.04514860
## DecisionTree        0.010680005 0.05220153 0.15607615 0.15174850 0.19014797
```

```
##                       Max. NA's
## LinearRegression    0.15529485    0
## nonLinearRegression 0.08976176    0
## DecisionTree        0.40569818    0
##
## Rsquared
##                       Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## LinearRegression       1       1      1    1       1    1    2
## nonLinearRegression    1       1      1    1       1    1    0
## DecisionTree          NA      NA     NA  NaN      NA   NA    9
```

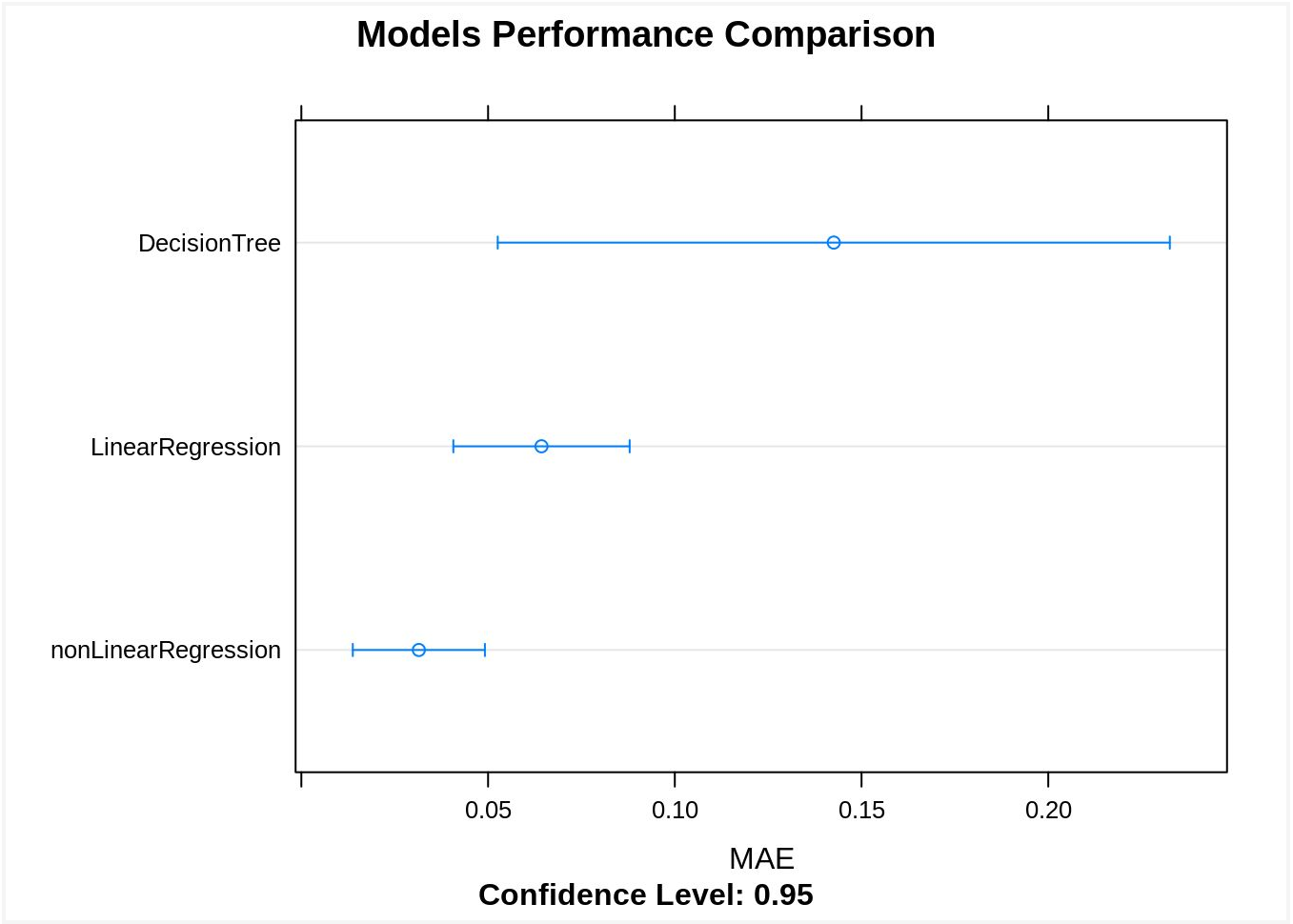


Figure 6: Models Performance Comparison.

Model selection:

The top performing model is the non linear regression one. We check whether the MAE gap between the top performing model and the 2nd best performing model pays off the complexity added by the non-linear regression. Simplicity is always to be favored if model performance is comparable.

Table 4 presents the results of comparing the MAE and the respective variance of the two models.

```
##
##  Welch Two Sample t-test
##
## data:  nlm_lift$resample$MAE and lm_lift$resample$MAE
## t = -2.5673, df = 14.831, p-value = 0.02159
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.060130333 -0.005548356
## sample estimates:
##  mean of x  mean of y
## 0.03146454 0.06430388
```

Table 4: Comparing the linear regression with the non-linear regression.

Based on the above reported results, the non linear regression model achieves a statistically significant better performance than the 2nd best model. (MAE = 0.031 and respective uncertainty 0.023). The added complexity improved significantly the performance and therefore this will be used to get predictions on the missing values originally requested.

For the reasons mentioned above, the team selected the non-linear model.

E. Model insights & Predictions

Besides picking the best performing model, drawing conclusions and insights is actually quite important as well:

Insights from summary results:

- It seems that customer quality is **not** important to predict the outcome
- There is evidence of a non linear relationship between discount and avg lift, verifying the respective plot in EDA section of this rept.
- Going back to the EDA figure 3, we can observe that the effect is more intense for discounts higher than 30%, identifying there a change point.
- In terms of future work: piece wise nature regression models could also be useful and should

be tested out in a next iteration.

Model Summary - Model weights and Variable importance

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.01715    0.01760  -0.975  0.35260
## discount            1.27142    0.33452   3.801  0.00348 **
## `I(discount^2)`    -4.46094    1.80236  -2.475  0.03282 *
## `I(discount^3)`     9.27491    2.42736   3.821  0.00337 **
## cust_quality.average  0.02745    0.01947   1.410  0.18899
## cust_quality.good     0.01368    0.02196   0.623  0.54716
## cust_quality.high     0.01660    0.02151   0.772  0.45812
## cust_quality.low      0.03090    0.02048   1.509  0.16229
## cust_quality.poor         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02529 on 10 degrees of freedom
## Multiple R-squared:  0.9884, Adjusted R-squared:  0.9804
## F-statistic: 122.2 on 7 and 10 DF,  p-value: 7.071e-09
```

Predictions

Using the non linear regression model we predict the missing values expecting a MAE according to the cross validation results already reported. Predictions can be seen in the following table:

	poor	low	average	good	high
50%	65%	69.4%	69.0%	67.6%	67.9%
40%	38%	40%	44%	38.5%	38.8%
30%	20%	26%	21%	19%	23%
20%	13%	16.4%	16%	18%	15%
10%	7.5%	10.6%	10.2%	8.8%	9%
0%	0%	0%	0%	0%	0%

Table 3: The quiz solution.

On Figure 7 we can see a visualisation of the quiz solution. We observe that discount has a positive impact on the surface which matches our expectation. The surface suffers from flex points that were mentioned earlier.

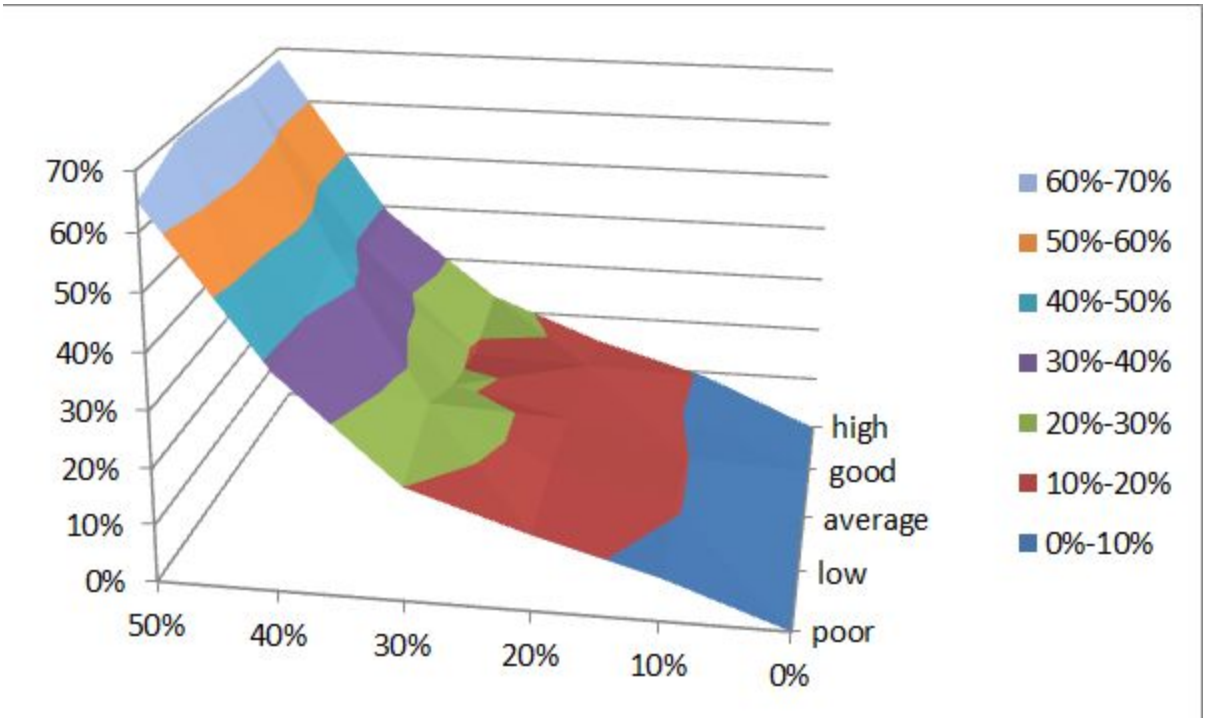


Figure 7: Visual presentation of the quiz solution.

On Figure 8 we see that the ensemble of original and predicted data stick together into groups and behave in a similar way. The customers' reaction to the 30% discount is the most dispersed group. The High group's response is sticky compared to the other customers' groups.

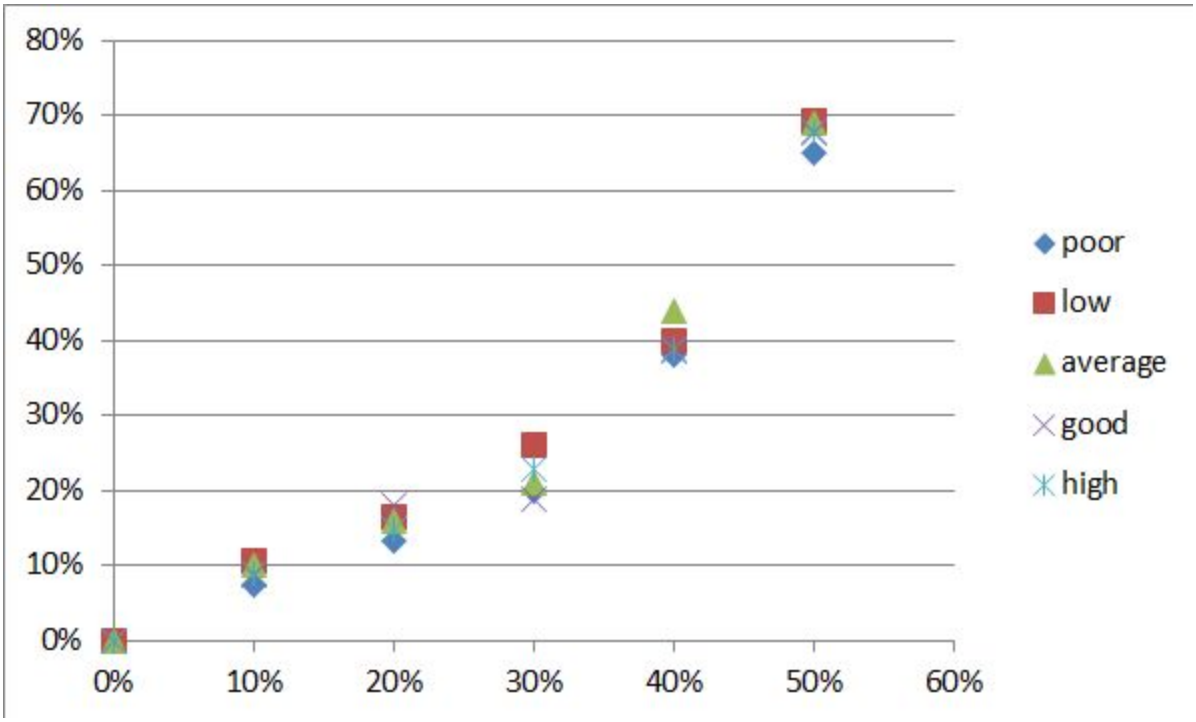


Figure 8: Clustering the quiz solution.

APPENDIX A

A bit more on the customer quality categorical variable and its relationship to the outcome avg lift. We will perform some hypothesis tests to investigate even further this relationship.

```
##
##  Welch Two Sample t-test
##
##  data:  train[train$cust_quality == "poor", ]$avg_lift and train[train$cust_quality == "low", ]$avg_lift
## t = 0.48321, df = 4.9926, p-value = 0.6494
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3781852  0.5531852
## sample estimates:
## mean of x mean of y
##    0.3075    0.2200
```

```
##
##  Welch Two Sample t-test
##
##  data:  train[train$cust_quality == "poor", ]$avg_lift and train[train$cust_quality == "average", ]$avg_lift
## t = 0.63515, df = 5.1917, p-value = 0.5523
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3152798  0.5252798
## sample estimates:
## mean of x mean of y
##    0.3075    0.2025
```

```
##
##  Welch Two Sample t-test
##
##  data:  train[train$cust_quality == "poor", ]$avg_lift and train[train$cust_quality == "good", ]$avg_lift
## t = 1.2179, df = 4.0754, p-value = 0.289
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2326389  0.6009722
## sample estimates:
## mean of x mean of y
```

```
## 0.3075000 0.1233333

##
##  Welch Two Sample t-test
##
##  data:  train[train$cust_quality == "poor", ]$avg_lift and train[train$cust_quality == "high", ]$avg_lift
## t = 1.2984, df = 3.7307, p-value = 0.2686
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2281085  0.6081085
## sample estimates:
## mean of x mean of y
##    0.3075    0.1175

##
##  Welch Two Sample t-test
##
##  data:  train[train$cust_quality == "low", ]$avg_lift and train[train$cust_quality == "average", ]$avg_lift
## t = 0.11797, df = 4.135, p-value = 0.9116
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3891223  0.4241223
## sample estimates:
## mean of x mean of y
##    0.2200    0.2025

##
##  Welch Two Sample t-test
##
##  data:  train[train$cust_quality == "low", ]$avg_lift and train[train$cust_quality == "good", ]$avg_lift
## t = 0.72981, df = 3.0307, p-value = 0.5178
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3224634  0.5157967
## sample estimates:
## mean of x mean of y
## 0.2200000 0.1233333
```

```
##
##  Welch Two Sample t-test
##
##  data:  train[train$cust_quality == "low", ]$avg_lift and train[train$cust_quality == "high", ]$avg_lift
## t = 0.80807, df = 2.6923, p-value = 0.4843
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3285716  0.5335716
## sample estimates:
## mean of x mean of y
##    0.2200    0.1175

##
##  Welch Two Sample t-test
##
##  data:  train[train$cust_quality == "average", ]$avg_lift and train[train$cust_quality == "good", ]$avg_lift
## t = 0.72017, df = 4.8552, p-value = 0.5046
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2059651  0.3642985
## sample estimates:
## mean of x mean of y
## 0.2025000 0.1233333

##
##  Welch Two Sample t-test
##
##  data:  train[train$cust_quality == "average", ]$avg_lift and train[train$cust_quality == "high", ]$avg_lift
## t = 0.82446, df = 4.5807, p-value = 0.4505
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1874876  0.3574876
## sample estimates:
## mean of x mean of y
##    0.2025    0.1175

##
##  Welch Two Sample t-test
```

```
##
## data:  train[train$cust_quality == "good", ]$avg_lift and train[train$cust_quality == "high", ]$avg_lift
## t = 0.074279, df = 4.1739, p-value = 0.9442
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2086717  0.2203384
## sample estimates:
## mean of x mean of y
## 0.1233333 0.1175000
```

Result:

Multiple hypothesis tests per group shows that no statistical significant differences are observed based on available samples (all reported p values are greater than 0.05 (p adjusted values as well). Hence the assertion that customer quality does not affect average lift is further enhanced.