Gasbarro Capstone Project: Opportunities Analysis - Machine Learning

August 2, 2019

Contents

Introduction

Back to Contents

In the previous assignments (see: OppAnalysis_DataStory_FINAL), we had explored various aspects of the dataset for sale opportunities. While some trends appeared to emerge (e.g. the number of WON opportunities seemed to decline over 3 years), there were outstanding factors (e.g. a company merger might have impacted sales data) and no strong correlation among the variables and observations.

As we embark on the linear modeling and machine learning phase of this project, we'd like to answer 2 key questions.

1. AGE. We had not used AGE in our previous data exploration and analysis. Age is the number of days between the start of the sale activity and the date when the opportunity is closed (WON, LOST). Is there a relationship or correlation between 1A. Age and how many deals are WON or LOST? 1B. Age and Expected Revenue?
2. LM and PREDICTION. 2A. If we know Industry, can we predict whether an opportunity will be WON or LOST, and how much the WON Expected Revenue might be? 2B. If we know Industry and Age, can we predict whether an opportunity will be WON or LOST, and how much the WON Expected Revenue might be?

The formula for prediction is based on the INPUT and the LM generated based on the variables in the dataset. The formula may be expressed as follows: *INPUT -> FUNCTION (LM Regression) -> OUTPUT Industry=Energy,Age=100 -> LM() -> Stage=?*

Sample Sales Data: Perfectly Prepared

Back to Contents

Before we use the real Opportunities data set, we wanted to see what the correlation and linear modeling would look like on a "perfect" data set, one in which there is a strong correlation between the variables data.

The Sample Sales Data has four variables: Industry, Stage, Amount, Days * All the Energy rows have stage=WON, Amount=250000, and Days=30 * All the Engineering rows have stage=WIP, Amount=100000, and Days=15 * All the Distribution rows have stage=LOST, Amount=0, and Days=6

Obviously, this is a perfectly clustered dataset which would not exist in the real world (or maybe it would, at the best software sales company). It gives a good example for us of what strong correlation, clustering, and LM would look like.

Is there a correlation between AGE and AMOUNT?

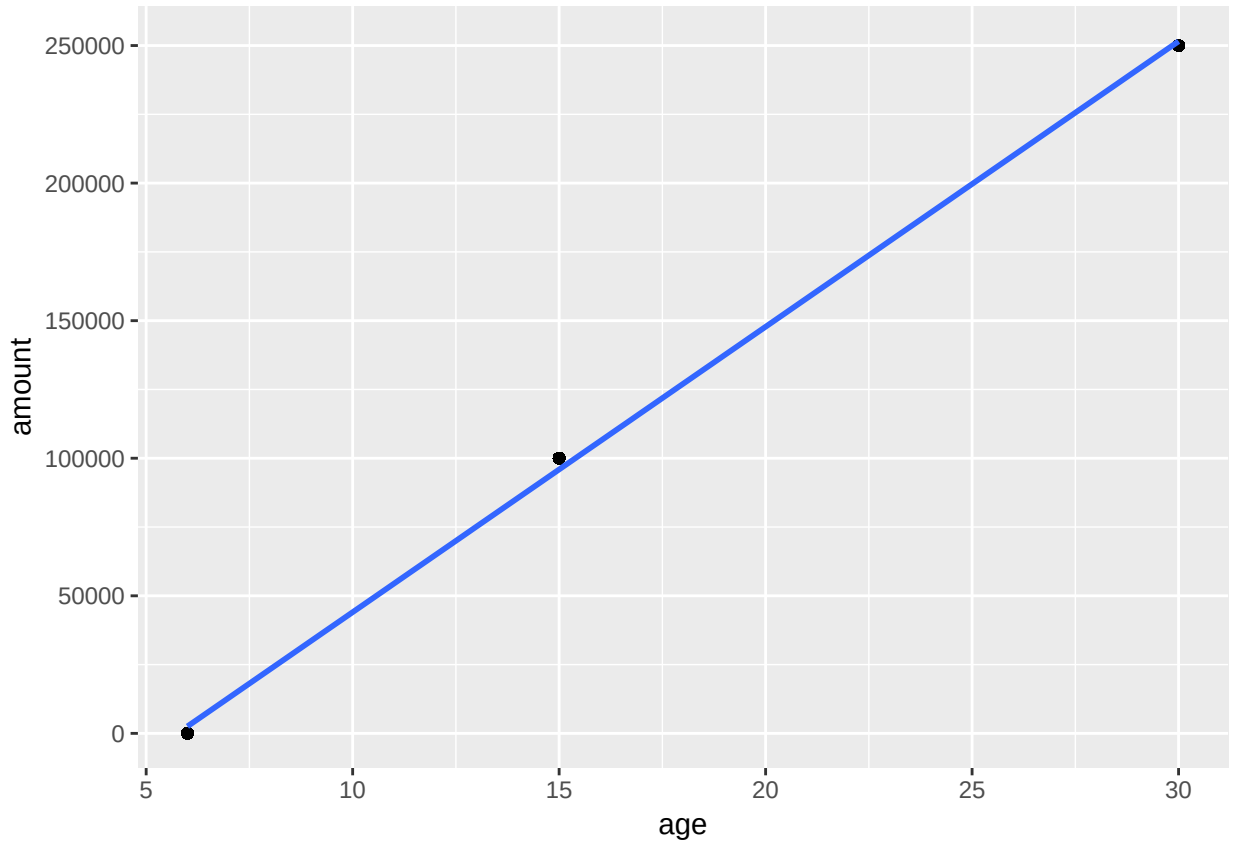
Back to Contents There should be. The number should be a very high correlation, quite close to 1.0 (or 100%).

```
cor(sample_sales_data1$age, sample_sales_data1$amount)
```

```
## [1] 0.9995971
```

When we plot this data in a scatterplot, the results shows a very clean, smooth line as expected.

```
ggplot(sample_sales_data1, aes(x=age, y=amount)) +
  geom_point() +
  geom_smooth(method="lm")
```



Linear Model for Sample Data

Back to Contents Lets build a linear model! This LM formula will be based on Amount against Age.

```
lm_amount <- lm(amount ~ age, data=sample_sales_data1)
summary(lm_amount)
```

```
##
## Call:
## lm(formula = amount ~ age, data = sample_sales_data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2551  -2551  -1531   4082   4082
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -59693.878    149.721  -398.7   <2e-16 ***
## age          10374.150      7.611  1363.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2918 on 1498 degrees of freedom
## Multiple R-squared:  0.9992, Adjusted R-squared:  0.9992
## F-statistic: 1.858e+06 on 1 and 1498 DF,  p-value: < 2.2e-16
```

```
mean(sample_sales_data1$amount) == mean(fitted.values(lm_amount))
```

```
## [1] TRUE
```

```
mean(residuals(lm_amount))
```

```
## [1] 2.081056e-12
```

Can we use the LM formula to predict an amount as a function of age? In the example below, we created a data frame that has Age=30. We would expect the predict result, based on the LM formula, to be around $250,000.

```
newsalesdata <- data.frame(age=30)
predict(lm_amount, newsalesdata)
```

```
##        1
## 251530.6
```

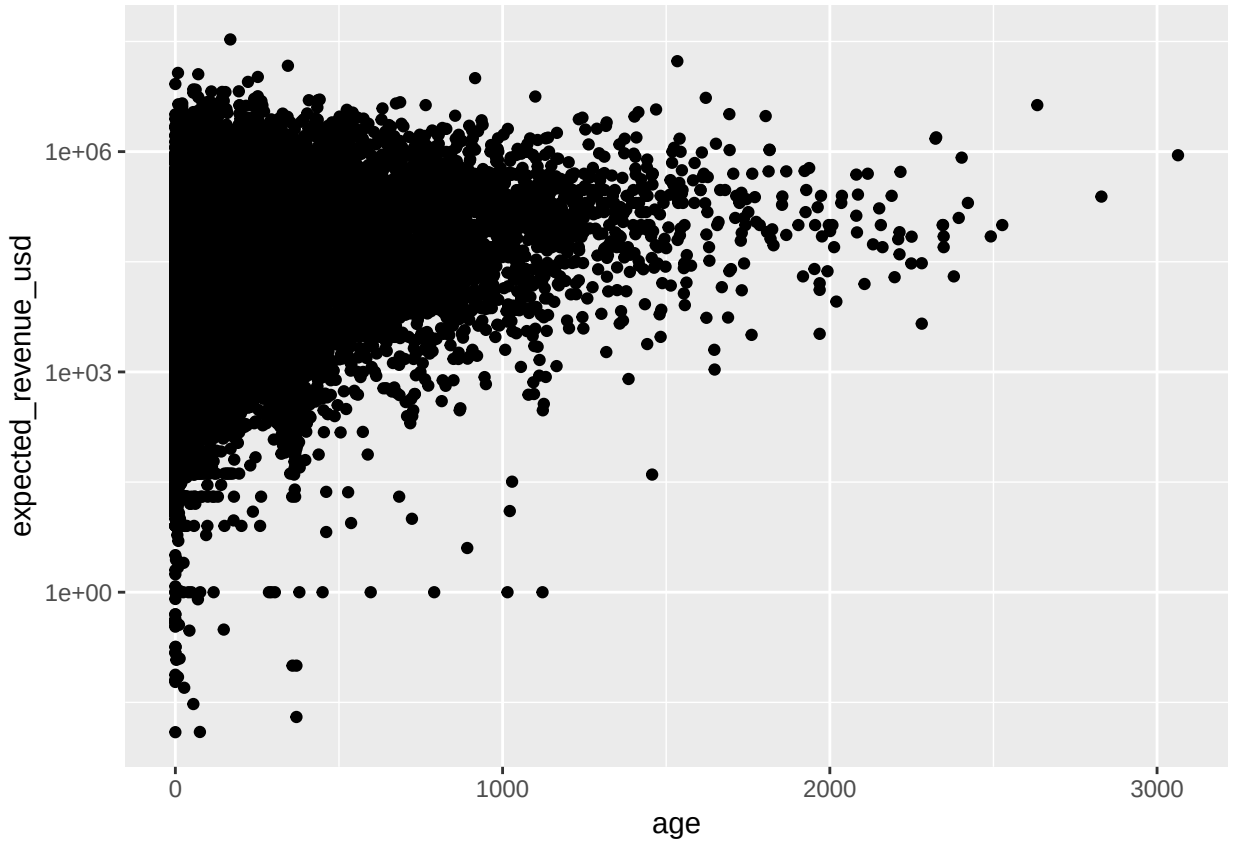
Opportunities: Correlation

Back to Contents Now let's work with the real sales data in Opportunities.

Is there a strong correlation between the AGE and the EXPECTED REVENUE among WON? The correlation is 0.1930255, which doesn't seem that significant.

What does the relationship look like when we plot using geom_point? It appears there is some correlation, as shown by the graph below.

```
ggplot(data = opps_won_2019, aes(x = age, y = expected_revenue_usd)) +
  geom_point() +
  scale_y_log10()
```



Opportunities: Linear Model

Back to Contents

We'll now create a linear model of the Expected Revenue as a function of Age. The Summar is printed below the code.

```
ageMod <- lm(expected_revenue_usd ~ age, data=opps_won_2019)
summary(ageMod)
```

```
##
## Call:
## lm(formula = expected_revenue_usd ~ age, data = opps_won_2019)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##  -666284   -32011   -28785   -18852 33611126
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29528.727   1013.879   29.12   <2e-16 ***
## age           276.064      4.931   55.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 262700 on 81000 degrees of freedom
## Multiple R-squared:  0.03726,    Adjusted R-squared:  0.03725
## F-statistic:  3135 on 1 and 81000 DF,  p-value: < 2.2e-16
```

```
ageMod_Intercept <- coef(ageMod)["(Intercept)"]
ageMod_Slope <- coef(ageMod)["age"]
```

The Intercept and slope numbers are interesting. While these numbers are large, neither of them seems to tell us whether this model is strong or weak.

The residual standard errors are quite high, indicating a lot of variance between the slope and the actual values in the data frame, suggesting that the linear model isn't very strong.

Finally, the R-squared values are barely .03 (on a scale of 0 to 1). Compared against the near-perfect linear model in our Sample Sales Set, which had an R-squared value of over .9, this seems to confirm that the model is not strong, nor will it's predictions be accurate.

But let's try anyway. :-)

Opportunities: Prediction

Back to Contents

We'll prepare a data frame with a new Age value, then run a prediction using the linear model ageMod generated in the previous exercise.

```
new_age <- data.frame(age=100)
new_age_predict <- predict(ageMod, newdata=new_age)
```

Based on the formula in the linear model, if the Age is 100 days, then the Expected Revenue for a WON deal is expected to be `5.713509\times 10^{4}`.

Thinking out loud: this model seems too simplistic. It is based solely on a single variable, Age. Other variables might impact the price in combination with Age, and offer a better model for making predictions.

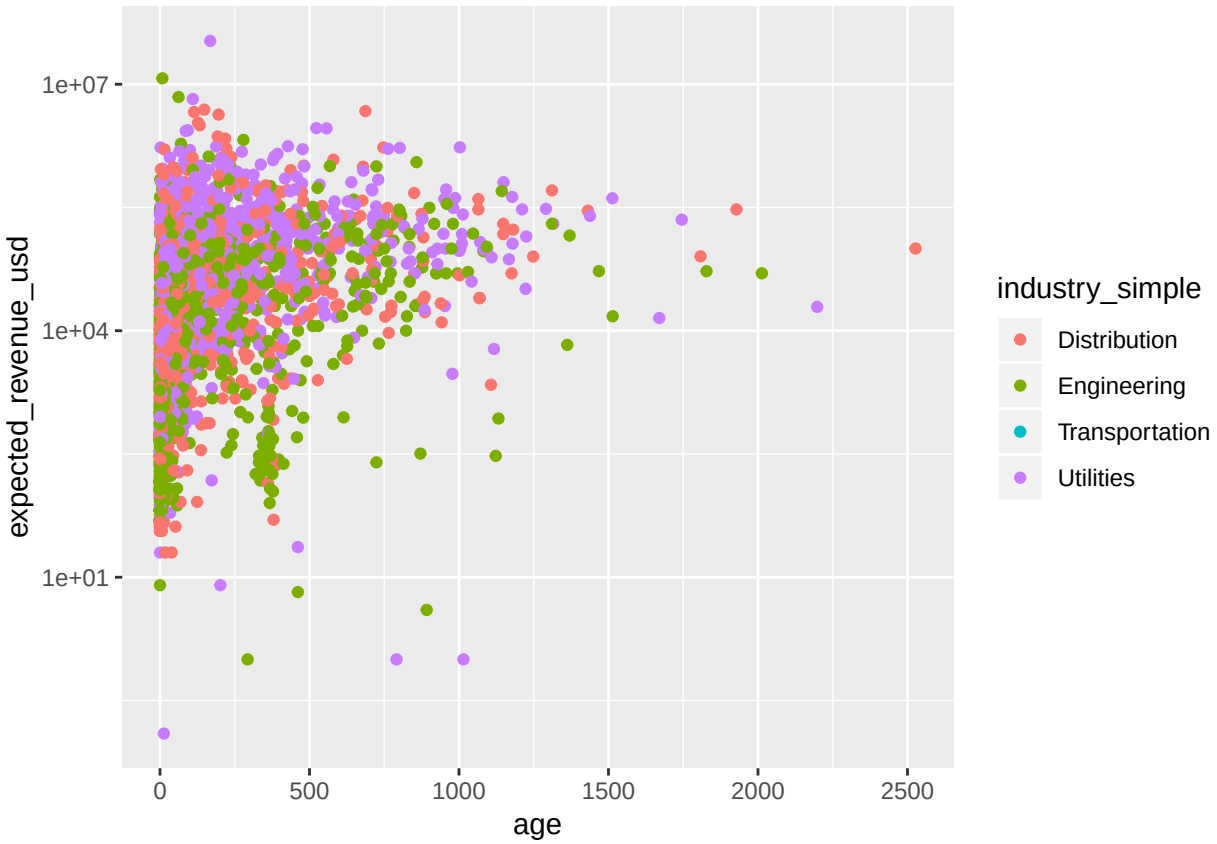
Opportunities: Multiple Regression

Back to Contents

Here is a scatterplot of Expected Revenue for WON opportunities, against Age and Industry Simple. In the first graph, the industry Everything Else dominates and masks the visbility of the other industries. In the second graph, Everything Else is excluded, giving us a better picture of the other industries.

```
ggplot(data = opps_won_2019, aes(x = age, y = expected_revenue_usd, col=industry_simple)) +
  geom_point() +
  scale_y_log10()
```


```
ggplot(data = subset(opps_won_2019,industry_simple %in% c("Transportation" , "Utilities", "Engineering"
  geom_point() +
  scale_y_log10()
```



We'll generate a linear model, adding Industry Simple as a factor to the original formula.

```
ageIndustryMod <- lm(expected_revenue_usd ~ age + factor(industry_simple), data=opps_won_2019)
summary(ageIndustryMod)
```

```
##
## Call:
## lm(formula = expected_revenue_usd ~ age + factor(industry_simple),
##     data = opps_won_2019)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##  -682123   -31614   -27566   -17833 33541275
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             48991.106   7086.417   6.913
## age                                       273.825      4.932  55.515
## factor(industry_simple)Engineering     -29533.474   9444.288  -3.127
## factor(industry_simple)Everything Else -20718.326   7132.238  -2.905
## factor(industry_simple)Transportation  -54005.006 262659.624  -0.206
## factor(industry_simple)Utilities        50764.559   9866.275   5.145
##                                         Pr(>|t|)
## (Intercept)                             4.77e-12 ***
## age                                      < 2e-16 ***
## factor(industry_simple)Engineering       0.00177 **
## factor(industry_simple)Everything Else   0.00367 **
## factor(industry_simple)Transportation    0.83710
```

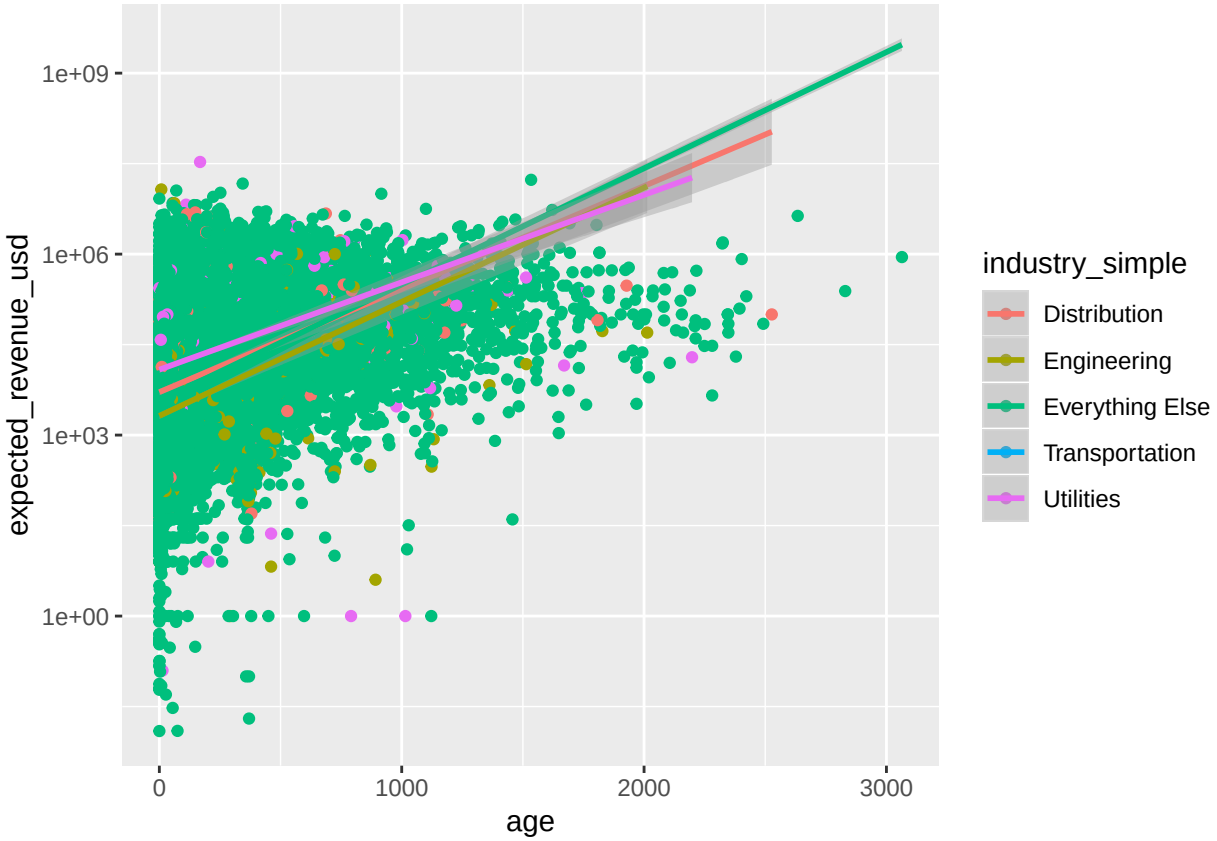
```
## factor(industry_simple)Utilities       2.68e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 262600 on 80996 degrees of freedom
## Multiple R-squared:  0.03863,    Adjusted R-squared:  0.03857
## F-statistic:   651 on 5 and 80996 DF,  p-value: < 2.2e-16
```

```
ageIndustryMod_Intercept <- coef(ageMod)["(Intercept)"]
ageIndustryMod_Slope <- coef(ageMod)["age"]
```

The Residual Standard Error seems in alignment with the previous linear model for Revenue and Age. What's interesting is the Multiple R-Squared has risen from 0.02 to almost 0.04. The number is still quite small, but this suggests that adding additional variables produces a model that will have more valid or accurate predict results.

The graphs below show the linear model applied against all Industries, then against the focused industries (excluding Everything Else).

```
ggplot(data = opps_won_2019, aes(x = age, y = expected_revenue_usd, col=industry_simple)) +
  geom_point() +
  scale_y_log10() +
  geom_smooth(method="lm")
```




```
ggplot(data = subset(opps_won_2019,industry_simple %in% c("Transportation" , "Utilities", "Engineering"
  geom_point() +
  scale_y_log10() +
  geom_smooth(method="lm")
```


expected_revenue_usd
1e+07
1e+04
1e+01
0
500
1000
1500
2000
2500
age
industry_simple
Distribution
Engineering
Transportation
Utilities

Conclusion: Significant or Unusual?

Back to Contents

We created a simple linear model for Expected Revenue and Age, then modified it to include a factor variable (Industry Simple). The initial results appear to show that these models are weak, would not be reliable for analysis or predicatability.

However, there appeared to be improvement when we added the factor Industry to the simpler linear model. This suggests that adding additional variables in some combination, or re-scoping and combining other variables (numericals like Forecasted Amount or factors like Stage and Currency) might produce models that are stronger, and cab eused to identify patterns, trends, and possible predictyions for future behavior.