Gasbarro Capstone Project: Opportunities Analysis - Machine Learning

July 16, 2019

Contents

Introduction **2**

Sample Sales Data: Perfectly Prepared **3**
Is there a correlation between AGE and AMOUNT? 4
Linear Model for Sample Data 5
Clustering the Sample Data 6

Opportunities: Correlation Between AGE and EXPECTED REVENUE? **7**

Opportunities: Correlation Between AGE and STAGE (WON or LOST)? **8**

Opportunities: Linear Model - Expected Revenue and Industry **9**

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?

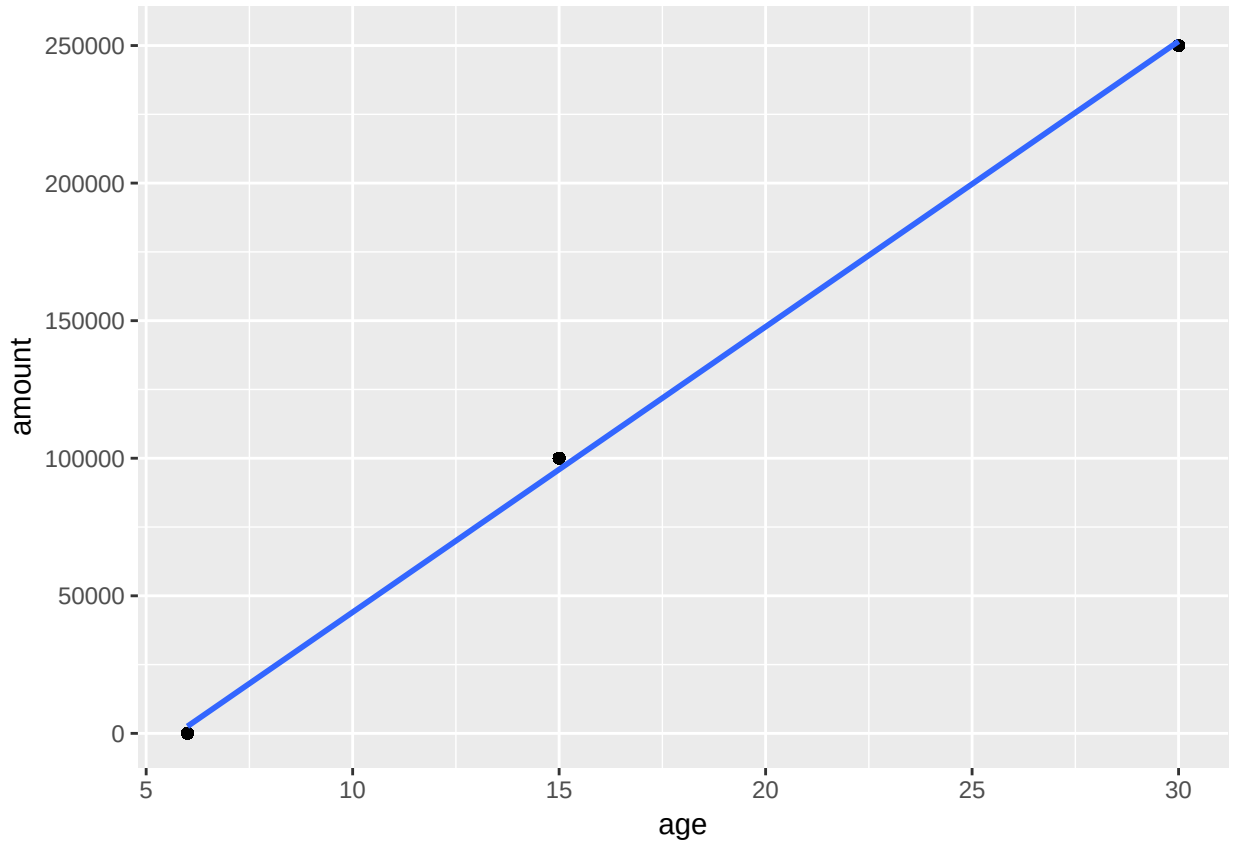
Back to Contents There should be. The number should be a very high correlation, quite close to 1.0 (or 100%).

```
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)
print(lm_amount)
```

```
##
## Call:
## lm(formula = amount ~ age, data = sample_sales_data1)
##
## Coefficients:
## (Intercept)          age
##      -59694        10374
```

Can we use the LM formula to predict an amount, based on the 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
```

QUESTION: What is considered a strong Intercept and/or value for LM?

Clustering the Sample Data

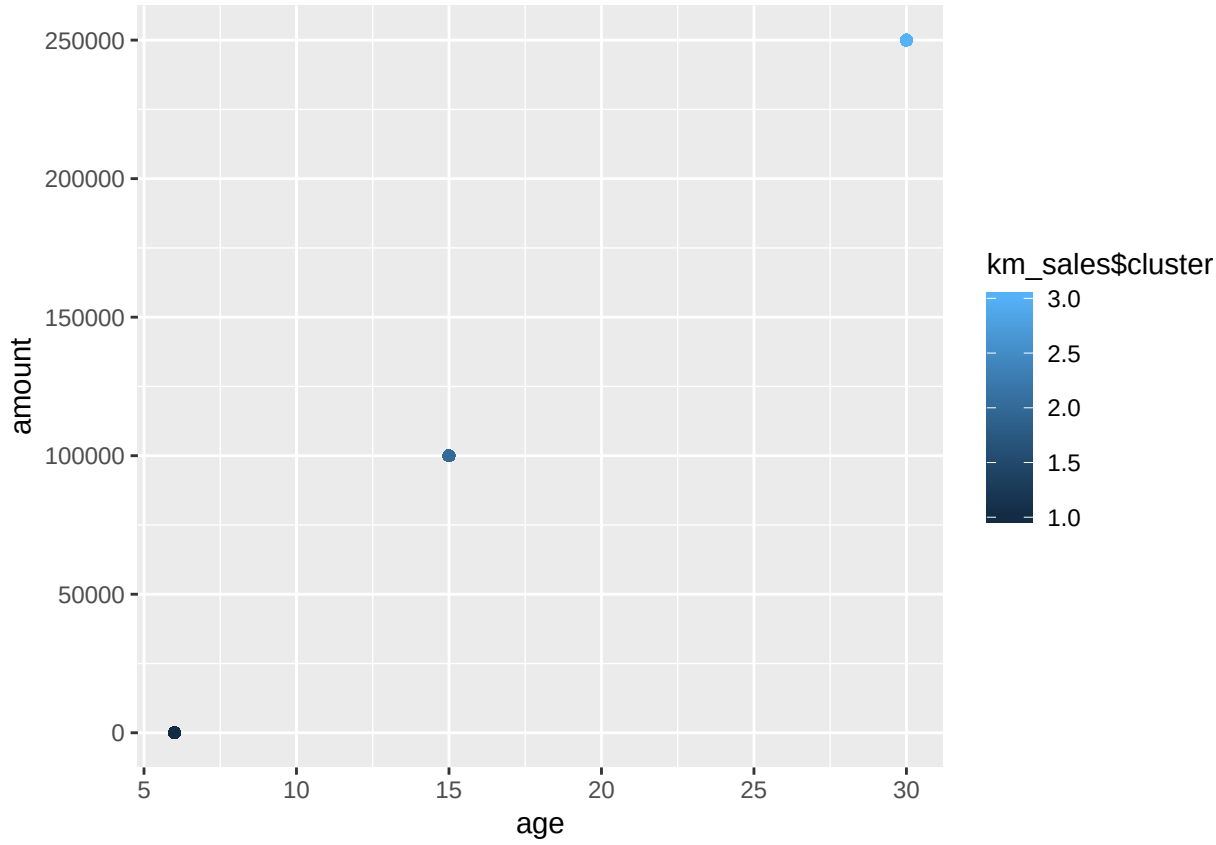
Back to Contents The sample data is intentionally clustered closely between all the variables. How would this look in a scatterplot? This gives us a perfect example to compare the actual Opportunities sales data against.

```
# First, separate the quantitative data from the qualitative
age_amount<-sample_sales_data1[3:4]
industry<-sample_sales_data1$industry
stage<-sample_sales_data1$stage

# Next, use kmeans to cluster the data into 3 parts (since we already know there are 3)
km_sales<-kmeans(age_amount, 3)
table(industry, km_sales$cluster)
```

```
##
## industry         1   2   3
##   Distribution 500   0   0
##   Energy         0   0 500
##   Engineering    0 500   0
```

```
ggplot(age_amount, aes(x=age,y=amount,col=km_sales$cluster)) +
  geom_point()
```



Opportunities: Correlation Between AGE and EXPECTED REVENUE?

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

Is there a strong correlation between the AGE and the EXPECTED REVENUE? Why do we see NA in this result?

```
cor(opps_revised_2019$age, opps_revised_2019$expected_revenue_usd)
```

```
## [1] NA
```

Opportunities: Correlation Between AGE and STAGE (WON or LOST)?

Back to Contents

We've been using correlation against purely quantitive data (age, amount). What if we want the correlation between quantitative data (age) and qualitative data(stage)? E.g. Is there a correlation between the Age of the opportunity, and whether the opportunity is WON or LOST?

Opportunities: Linear Model - Expected Revenue and Industry

Back to Contents

```
# Replace all NA values in industry_simple
#is.na(opps_revised_2019$industry_simple)
opps_revised_2019$industry_simple[is.na(opps_revised_2019$industry_simple)] <- "Everything Else"
#is.na(opps_revised_2019$stage_simple)

# Simple Linear Model Forecasted Amount by Industry
lm_forecasted_industry <- lm(forecasted_amount_usd ~ industry_simple, data=opps_revised_2019)
lm_forecasted_industry
```

```
##
## Call:
## lm(formula = forecasted_amount_usd ~ industry_simple, data = opps_revised_2019)
##
## Coefficients:
##                     (Intercept)      industry_simpleEngineering
##                          180421                          256944
## industry_simpleEverything Else   industry_simpleTransportation
##                          -57115                         -126056
##       industry_simpleUtilities
##                          -25689
```

```
#unseen01 <- data.frame(industry_simple="Energy")

#predict(lm_forecasted_industry, unseen01)
```