Springboard’s Foundations of Data Science

Stephanie Vaul

Capstone Project Proposal

1. What is the problem you want to solve?
2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?
3. What data are you going to use for this? How will you acquire this data?
4. In brief, outline your approach to solving this problem (knowing that this might change later).
5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

I want to analyze the BMW automotive sales trends in the USA over the last 10 years in order to determine which areas of the country and which model series display the highest sales volume and repeat ownership (brand loyalty). In addition, I will try to identify first-time BMW owners and see which model appeal most to them. I will also look at how these trends vary between New and Used vehicles.

My client is BMW North America, a company that my employer has partnered with over the last 10 years to provide customer retention marketing. Determining which model series do well in each geographic area, as well as which series result in repeat buyers will help BMW better focus their marketing efforts for various ad campaigns by leveraging existing trends, as well as being able to extrapolate those trends in other areas of similar makeup.

I plan to use dealership sales data that my company has been collecting for the past 10+ years in order to provide sales to service retention marketing, e.g. initial service reminders. This data includes customer and vehicle information, along with the purchase date and type (new vs used). Some of the data includes demographic information, like birthdate and gender. I will review the data to see if there is sufficient data in order to provide trend reports based on age &/or gender.

Proprietary sales data from most BMW centers in the U.S. going back 10yrs

**PurchDate**

**VehYear**

**VehModel**

**TradeVIN**

**TradeVehYear**

**TradeVehModel**

**NewUsed** (New, Used, Other sale)

**PurchType** (Lease, Cash, Finance)

**CustState**

**CustZip**

**CustID** – to look for repeat purchases

**Gender** – if enough % of records are populated

**Birthdate** – if enough % of records are populated

BMW also has their own designations for dealership Region and Market. I will apply this similar sorting parameters to the customer’s address in order to determine which Region a customer belongs too.

Because we also collect service data, I will be able to also look at how long a customer owned a vehicle by reviewing both Sales and Service data. I will first look to see if we have subsequent sales data on the vehicle – whether that be from another customer purchasing the VIN or the VIN being listed by the same customer as being traded in. If there is no additional sales record, then I can turn to the service history to see when they last serviced it. While this is not entirely accurate should the customer service at an independent station, it will provide better information than not having it at all.

My approach will be to first segment the customers into the various BMW regions, possibly down to the Market level. Their markets are based on BMW locations so this might not be to be extrapolated correctly to all customers. I will then segment out the purchases by model series and type (new or used). I want to be able to see a count of each model series and type for each year overall and within each region. Once top-level exploratory analysis has been performed, I’ll start looking into more details on customer loyalty based on repeat purchases and try to identify which are made by first-time buyers. After that, I will look to see if I can pinpoint the length of ownership for each vehicle to be able to summarize which series are kept by a single owner the longest.

I will provide the code snippets used to analyze these trends, along with a paper of my findings and a presentation slide deck.