**Milestone Report – 11/21/2016 Andres Forero**

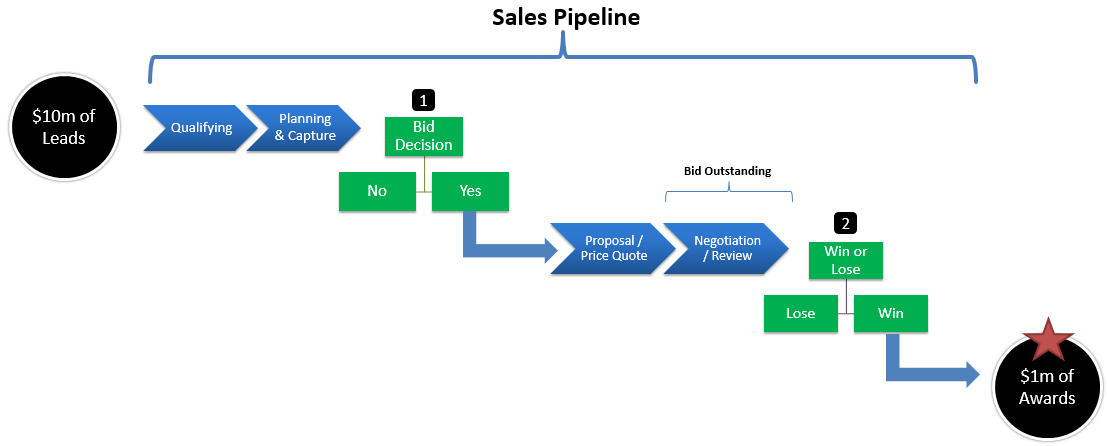
**Capstone Project: Using the Sales Pipeline to predict the $ amount of “Awards” won.**

**1. What is the problem you want to solve?**

* My company uses Salesforce to manage its sales pipeline. An ongoing issue is knowing whether our pipeline is large enough, and generates enough “Award” wins to sustain future revenue growth targets.
* I want to use a machine learning categorization algorithm to forecast the win / lose outcomes of our sales pipeline leads.

**Example**:

The diagram pictured below is a visual representation of my company’s sales pipeline. Using a made up number, suppose that $10m in leads (or opportunities) enter the pipeline. The leads will progress through the “Qualifying” and the “Planning & Capture” stages, whereby proposal analysts collect data on those leads. Then Decision 1 – the Bid Decision will be made, whereby the opportunity will either be Bid on or not. The $ amounts that are “NO” for BID will fall off the pipeline. Those that are “YES” for BID will progress through the next stages of the pipeline. At the end of the Proposal / Price Quote stage, whereby a proposal is submitted to the client, the opportunity is said to be “Outstanding”, meaning we are waiting for the client to provide feedback for “Negotiation / Review”. Finally, Decision 2 – the Win or Lose Decision will be made. If we lose the bid, the $ amount falls out of the pipeline, while if we win the bid, the $ amount of the opportunity becomes an “Award”. In hypothetical example pictured in diagram, we started with $10m in opportunities or leads, and ended up with $1m in “Awards” or opportunities won. What I would like to predict better using Machine Learning is the $ amount at the end of the pipeline diagram.



**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?**

The client is the CFO, and he cares about this problem because:

* Knowing the “Award” wins allows him to make sure that our sales pipeline can generate enough revenue to surpass future growth targets for his 3-year strategic plan.

Based on my analysis, what will the CFO do that they wouldn’t have otherwise?

* Supposing that my analysis forecasts wins more accurately than the current model, and it shows that we are not expected to reach growth targets, the CFO can make decisions that get the institute back on track.

**3. What important fields and information does the data set have?**

These are the most important columns (there are more columns than these, but in interest of not commenting on every single column, I just picked the most important ones):

1. **Bid\_Recommendation\_Bid\_Status** = this column has the following possible values:
   * + Bid
     + No Bid
     + Cancelled
     + Blank (i.e. no recommendation entered in system)
2. **Stage** = this column has the following possible pipeline stages:
   * Closed Won
   * Closed Lost
   * Closed Other
   * Proposal / Price Quote
   * Negotiation
   * Review
3. **Bid Probability %** = this column has possible values of 50% or 100%, which are entered by the project leaders. Some rows contain zero values, and so that data needs to be cleaned up with code (described later in this document).
4. **Amount =** this columnshowsthe expected $ amount of the contract as moving through the pipeline stages.
5. **Amount Range** = based on the ‘Amount’ column, and has these possible $ ranges:
   * 1 - 1m
   * 1m – 10m
   * 10m – 20m
   * 30m +
6. **Probability of Winning %** = this is a column that is populated by the project leaders and has these possible values:
   * 10%
   * 30%
   * 50%
   * 70%
   * 90%
   * 100%
7. **Estimated Fee %** = this column shows the estimated fee % expected to be earned on the contract.
8. **EstimatedB&PAmt** = this column shows the estimated $ amount that the Business & Proposal team will spend on bidding on the opportunity (lead). One hypothesis is that the higher the amount of the opportunity, the higher the B&P amount.
9. **AdjEstimatedBase+OptionDuration** = this columns shows the number of days that an opportunity (lead) is expected to last if the contract is won.
10. **Pipeline Stage History (Days)** = this section is broken up into the following columns, and each contains the number of days that the opportunity spent in each of the pipeline stages.
    * Days in Qualitfication
    * Days\_in\_Planning&Capture
    * Days\_in\_Proposal/PriceQuote
    * Days\_in\_Negotiation
    * Days\_in\_Review
11. **Market** = this column is the market that the opportunity belongs to and contains the following possible categorical values:
    * Education
    * Energy & Environment
    * Food & Agriculture
    * Health
    * Multi-area
    * Additional Markets

**4. What are its limitations i.e. what are some questions that you cannot answer with this data set?**

In general, the limitations of the data are based on how accurate the data is in itself. For example, what process is being used to populate the “Probability of Winning %” field. Are people just guesstimating out of the air, or is there a more rigorous process to come up with the probability? The accuracy and process by which fields are populated in Salesforce will have a great effect on how accurate the prediction model will be. In addition, in trying to predict win or losses of contracts, it would have been useful to have the following columns or data:

* # of Competitors also bidding on contract
* Name and information of specific competitors
* Comments by Project Leader on likelihood of winning
* Win rate of internal sales team bidding on the contract

This data is currently not being populated in Salesforce, therefore, the robustness of the model will be limited by the data that is currently available.

Another question I cannot answer, is why we decided to NOT BID on certain opportunities. What I can tell you is that for these opportunities, the Probability of Winning was in the 30% range. However, besides “Probability of Winning” being lower than 35%, there is no other field that is common among these opportunities that we did NOT BID on.

**5. What kind of cleaning and wrangling did you need to do?**

* The “**Amount**” column contains several zero values, so those need to be excluded from the final data frame.
* I created several binary columns (possible values are either 1 or 0), such as “**Bid\_NoBid**”, “**Win\_Lose**”, or “**Probability of Winning**”, which are based on other already existing categorical columns. These columns should be useful when implementing the machine learning algorithm.
* **BidProbability %** -> several rows contain zero or “undetermined” values for this column. To clean up, any rows where the opportunity has made it through the “Bid / No Bid” Decision (in other words are in the “Proposal / Price Quote” or the “Negotiation / Review” stages), automatically get populated with 100% bid probability.
* “**AdjEstimatedBase+OptionDuration**” -> several rows contain zero values for this column. To clean up, the field will be populated based on the following 4 conditions: **1)** If “Amount Range” = 1 – 1m, then 15, **2)** If “Amount Range” = 1m – 10m, then 30, **3)** If “Amount Range” = 10m – 30m, then 48, **4)** If “Amount Range” = 30m >, then 60

**6. Are there other datasets you can find, use and combine with, to answer the questions that matter?**

No, unfortunately, there are no other datasets which could be used.

**7. Any preliminary exploration you’ve performed and your initial findings. Test the hypotheses one at a time. Often, the data story emerges as a result of a sequence of testing hypothesis e.g. you first tested if X was true, and because it wasn't, you tried Y, which turned out to be true.**

Yes, I performed several hypothesis tests, which you can view in the Jupyter Notebook called “**Capstone Milestone Report.ipynb”**

**8. Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?**

My approach has not changed. I still want to use a machine learning categorization algorithm to forecast the win / lose outcomes of the sales pipeline leads. The hypothesis tests I performed provide some interesting insights, but I do not think that they lead to a change in my initial approach.