Elizabeth Siami

12/21/16

**Project Design Writeup and Approval Template**

**Problem to solve**

Based upon factors (listed below), determine the likelihood of a current NetBase customer churning, renewing or growing (upsell) during their next renewal cycle. This will help our services team forecast revenue with higher accuracy.

**Churn** – if a customer cancels their contract, ACV is $0.

**Renewal** – if a customer renews for the same ACV (or less). No change to existing contract.

**Upsell** – if a customer renews and increases total spend (ACV is higher)

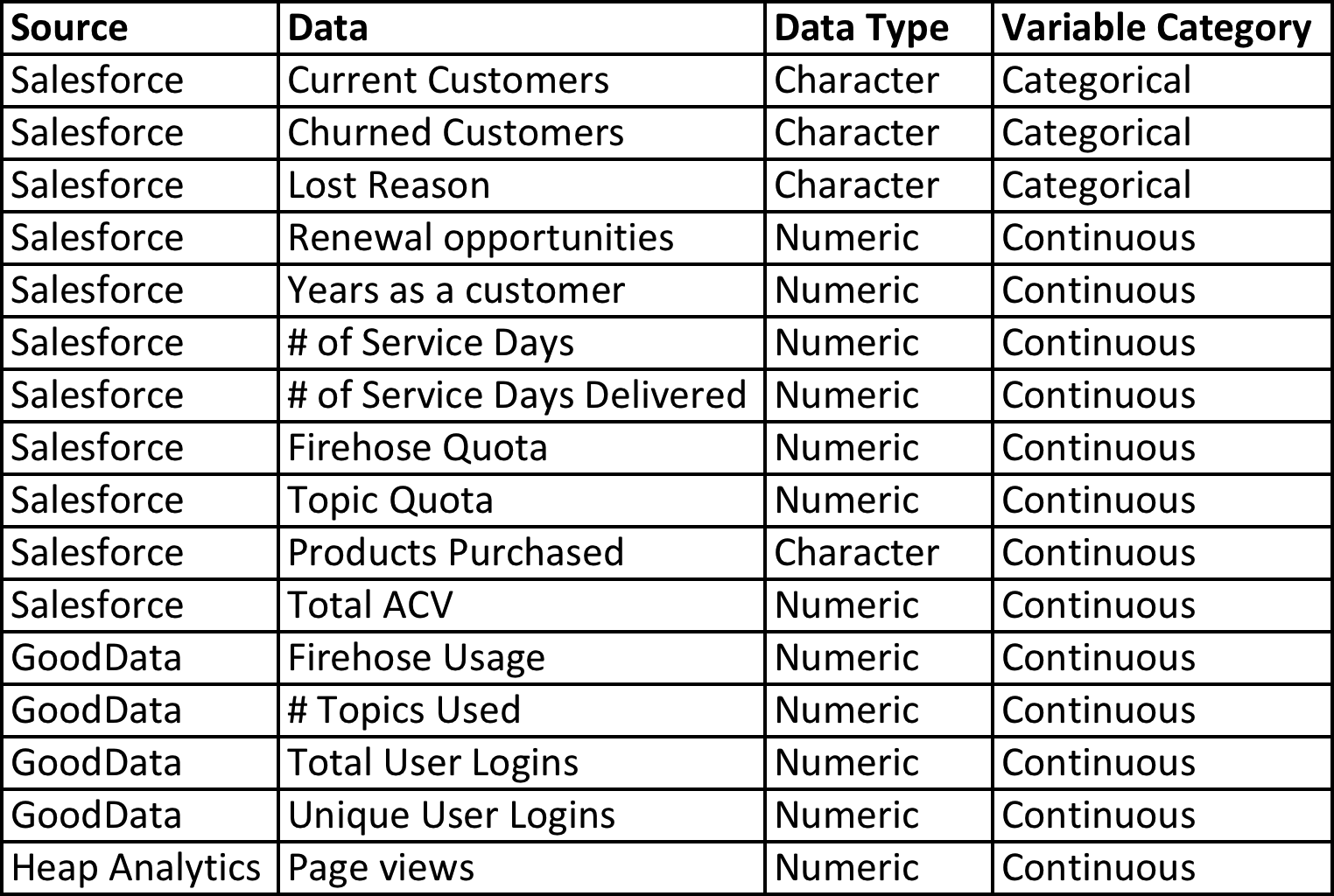
This model will predict a binary value – will the customer churn, renew or upsell? Therefore, this will be a classification model.

Our renewal rate is strong – the annual average is 85%, so what will make the most impact is avoiding churn, thus bringing up the renewal rate and increasing the accounts that upsell. This will set us on the path to profitability.

**Datasets**

The data will come from our Salesforce, GoodData and Heap Analytics instances. This data is available at the opportunity level, so I can use all associated renewal information to predict churn or upsell likelihood.

Below is more information on the datasets – the source, field, type and variable category.



**Domain knowledge**

I have previously worked on Client Services teams at two companies, so have worked closely with customers – building relationships, discussing renewals, upsells and work to prevent churn. I have been an analyst for 2+ years, and support the post sales teams and work closely with our customer data. We implemented a tool called Gainsight for our post sales teams to track customer usage, health and ultimately their chance of renewal. Because the tool is only valuable based upon the manual work/updates from the services teams, the predictor ability is poor. My goal is to build a model that uses various data points to determine if an account will renew (or churn) and the probability of an upsell without any manual effort.

I have yet to see any models that predict customer churn. What you can find on Google is more anecdotal – different company experiences and thoughts on early indicators of churn or upsell. The factors and weighting included in a prediction also vary tremendously by company. I also spoke with the Data Science team at my company, and they have only built models using social media data, not customer usage. So, this would be the first time someone at NetBase has built a model to determine the likelihood of retention based upon customer usage data.

**Project Concerns**

Below are a few of my concerns with the project:

* **What type of analysis to use?** Do I need to run three different models to determine correlation between the variables and churn, renewal and upsell separately? And then run another model to compare the likelihood of churn, renewal and upsell based upon the correlation? How complex does the model need to be? Is this a classification problem or will a linear regression suffice?
* **Data aggregation** - I am unsure how to aggregate the data, namely the FH usage, user logins and page views. This data can be pulled daily, weekly, monthly etc. but do not know the best way to aggregate the data so that it makes the most sense for the model. Should I average the monthly usage, compare that to the quota to determine % usage? Same with login data – should I average the monthly logins, and compare it to the # of licenses purchased?
* **Small dataset** - Since my dataset is small (~350 existing customers, ~50 churned customers), it may skew the results and therefore inflate or deflate the likelihood of one of the three outcomes. As we continue to sign up more customers, and lose more customers (hopefully not), I assume the predictor will become more and more accurate.
* **Do I need to convert my data**? I am unclear if I need to convert my data into dummy variables, or if any other data conversions are necessary

The cost of the model being wrong is incorrect prediction of customer churn and/or renewal that significantly impacts our recurring revenue, forecasting and headcount/hiring plan. If we can better predict retention and upsell, we can more accurately forecast not only revenue, but if we’ll need to hire more people on the services team to support the growing customer base.

Some of the data could be incorrect. A lot of the data coming from Salesforce is manually input by our Sales Operations team, so there can be mistakes from time to time (like the # of Services Days or the Firehose Quota).

**Outcomes**

I expect the model to show three different classes – one for customers that are predicted to churn during their next renewal cycle, customers that will renew, and customers that will upsell. I also would expect that the 3 classes will not be clearly distinct, and that some customers may fall into more than one class – like maybe based on the data, a customer would have a 50/50 probability of churning and renewing.

Since I have a limited dataset, I don’t think my model will able to classify the outcome with 100% accuracy. So, if we can accurately predict 50%+ of the outcomes, that would be a success. As I mentioned earlier, the more data we add to the model, the more accurate the prediction will be.