**Project Proposal**

**As of April 26, 2018**

1. Name, Date.
2. High level description of the project: what question or problem are you addressing?
3. How are you presenting your work? (web app, visualization, presentation/slides, etc.)
4. Whats your next step?
5. Data

* link to and description of data source
* please note that proposals without specific data source cannot be considered

All of my projects will use GitHub as a wrapper and will be presented as slides.

**Project 1: Slow Moving Inventory**

Slow moving inventory is hard to forecast and is traditionally modeled using the poisson distribution. Depending on the cost of the merchandise and the company inventory holding costs, it can be really expensive to stock inventory and even harder to offload.  
What if we could do a better job forecasting using another parameter or modeling the data with a nonparametric learner?   
Regression

**Dataset:**

* It has about 198K items. Not a lot of dimensions and not much cleaning to do.
* There are two variables that can be used as targets: binary (sold in the past 6 mos or not) for classification and SoldCount (how many items were sold) for regression.
* The x variables are limited: there’s MarketingType (how product was marketed), whether it’s a new product or not, release year (which is not really a knob we can turn, but it may be worth exploring), 3 price categories (PriceReg, LowUserPrice, LowNetPrice – I am thinking they are regular price, discounted price, and net cost of the product) plus a thing called StrengthFactor (whatever that is – it’s a very large number).
* There’s also an ItemCount, which I think is accumulated unsold inventory. So you could calculate how much the accumulated inventory is costing the company now and how much each percent improvement in forecasting would mean.
* Also there’s an order number and SKU number, so products can be aggregated by SKU.

Next Steps:

* EDA – see if I can generate some pretty charts to liven up the slides
* Determining which models to use and prepping the data accordingly (scale/standardize, treat null values, binary variables, categorical variables, etc.)

**Project 2: Converting Churned Customers to Active with Channel Profit Maximization**

Got idea from case study  
Use probabilities for profit curves and come up with channel coordinating profit proposal for retailer & producer  
Classification   
Needs dataset - fictional profit figures

[http://archive.ics.uci.edu/ml/datasets/Online+Retail#](http://archive.ics.uci.edu/ml/datasets/Online+Retail)

**Dataset:**

* It has over 541K items. Not a lot of dimensions (8) and not much cleaning to do.
* There are two variables that can be used as targets: binary (sold in the past 6 mos or not) for classification and SoldCount (how many items were sold) for regression.
* The x variables are limited: there’s MarketingType (how product was marketed), whether it’s a new product

**Project 3: Forecasting Beer Demand and Optimizing Production Planning**

This idea comes from an amazing dataset on Kaggle on a fictional beer wholesaler. You are basically predicting two things: first, beer demand for January 2018 based on past demand using promotional data, pricing, weather data, soda sales, sporting events and other beer-worthy occasions, and customer demographics. Second, you need to make beer product-mix recommendations for two new retailers that are joining your company as clients.

Obviously, this is great for regression but also for time series analysis (which has lots of awesome algorithms for smoothing and trends and seasonality and whatnot). Also, as a second—prescriptive analytics—step, one could forecast for the whole year and use fictional numbers to create either an optimized master production plan or optimize the distribution network now that there are two new customers.

<https://www.kaggle.com/utathya/future-volume-prediction>

**Dataset:**

* It has several csv’s (see list below). It has lots of features but only 21K instances. There’s absolutely no cleaning.
* price\_sales\_promotion.csv: ($/hectoliter) Holds the price, sales & promotion in dollar value per hectoliter at Agency-SKU-month level
* historical\_volume.csv: (hectoliters) Holds sales data at Agency-SKU-month level from Jan 2013 to Dec 2017
* weather.csv: (Degree Celsius) Holds average maximum temperature at Agency-month level
* industry\_soda\_sales.csv: (hectoliters) Holds industry level soda sales
* event\_calendar.csv: Holds event details (sports, carnivals, etc.)
* industry\_volume.csv: (hectoliters) Holds industry actual beer volume
* demographics.csv: Holds demographic details (Yearly income in $)

Project 4: Candy   
4. Forecasting given long lead times due to ocean shipping  
Time series (lots of algorithms)  
Ties into inventory  
Kaggle two datasets (one straight weekly forecast, other combined with 4 warehouses)  
  
3. Transportation approaches  
Wholesale customer dataset on UCI  
Fresh, grocery, frozen, etc.  
Is it cheaper to operate ambient trucks or dedicated fleet of trucks? Where’s the decision line? What are the driving variables to watch?  
Fictional costs  
3 Portuguese cities  
  
5. Production optimization  
Needs dataset (maybe candy demand dataset from kaggle? Inputs would be totally fictional)  
Straight up optimization - not very technically impressive

**Project 4: Forecasting Seasonal Candy Demand for the Year and Optimizing Production Planning**

I am a bit hesitant about this project because it comes from Kaggle and it already has some good work on it. Basically, candy consumption is extremely seasonal and this dataset shows production (sadly not demand) of   
  
Needs dataset (maybe candy demand dataset from kaggle? Inputs would be totally fictional)  
Straight up optimization - not very technically impressive

<https://www.kaggle.com/rtatman/us-candy-production-by-month>

Look at visualizations!!! Very heloful!!!