CAPSTONE PROPOSAL : Banking the Marijuana Business

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

Financial services companies (banks, credit unions, community banks, private lenders) assess the risk and reward of doing business with clients during their KYC (Know Your Customer) process. We will help solve the problem of assessing the risk and reward of doing business with their new target customer segment of marijuana businesses.

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?**

Financial services companies, traditionally, are very conservative, yet they constantly trade-off the risk and reward of any decision. The Banking client will decide whether to do business with or NOT do business with (call de-market) marijuana business owners.

3. **What data are you going to use for this? How will you acquire this data?**

I will use public data sources acquired from the Washington Cannabis Board and Secretary of State:

a. WA Secretary of State - Business registration data including trade name, address, owners, etc
b. Cannabis board - Companies who have applied for licenses including the state of the approval (pending, approved, etc) and longitudinal sales and violations (for those approved businesses).

4. **In brief, outline your approach to solving this problem (knowing that this might change later).**

I will first wrangle the data into clean forms for analysis. This consists of "tidying" the data (see Tidy in R) for 3 - 4 sets of data in CSV format. I will look for missing, incomplete, or clear mistakes in data and note any updates in my script in order to yield reproducible research.

Next, I will explore the data likely using Tableau, Python and R. I will look for areas of further exploration in statistical inference and prediction identifying the most interesting set of business questions to answer.

5. **What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.**

I will produce a set of scripts, a Tableau story, and an online form of the story (interactive blog post or slideshare deck).

********** PREVIOUS IDEAS **********

1) Acquiring news signal for customer risk assessment (NLP)

- Problem - Develop a customer risk rating model based on negative news signal acquired from available sources (news, social media, public record)
- Enabled decision - Whether to continue banking relationship or de-market banking customer
- Data available - https://www.dropbox.com/sh/gy350oj56u4hvbl/AAD8d1-b3Hye13I9UVLTYIQwa?dl=0
- Approach
    - Acquire a corpus of data on watch lists (marijuana-related business and special designated nationals) and related data from open sources (watch lists themselves, OpenCorporates, Google search results, LittleSis)
    - Detect entities, relations, keywords, topics, sentiment, and relevance from articles/text

    - Model into graph to uncover interesting relationships between people and companies
    - Use graph analytics in Spark/GraphX including (PageRank, connected components, label propagation, SVD++, Strongly connected components, triangle count)
    - CONSIDER: What can/should be shared (data, techniques, etc)
- Deliverables
    - Cleansed dataset for building the graph to connect entities (people/companies) with linking fields (geography, name)
    - Descriptive statistics to facilitate exploratory analysis
    - Graph model creating triangles or communities
    - Visualization illustrating factors of risk (connectedness to high risk/negative sentiment)
    - Report out of findings and recommended applications

NLP PRACTICE Dataset: Hilary Clinton Email - Use Data Story (Tableau Public)
https://www.kaggle.com/c/hillary-clinton-emails/data

2) Real estate value prediction/heat map (need to scope better and get rich data)

- Problem - Young (millennials) or down-sizing (pre-senior) couples looking for a 2-person home (to rent, own or rent to someone) are seeking housing value. In a growth market like Seattle, where are the neighborhoods with strongest potential to hold value for both renters and buyers over the next 5 years?
- Enabled decision - Choose a neighborhood that holds value
- Data available - https://www.dropbox.com/sh/pbgm96o64ygzih1/AAAOlBGsARdNw-L923bxcDbHa?dl=0
- Approach
    - Define factors of a “value” neighborhood
    - Collect data describing those factors (in Seattle)
    - Describe current status of Seattle neighborhood using current status of factors
    - Choose a model (classification or regression) to classify Seattle neighborhoods future in 5 years
- Deliverables
    - Cleansed dataset
    - Exploratory data analysis (tables, clorox pleths) - draw similarities between needs of millenials and downsizing pre-seniors assuming diverse neighborhoods are good for both demographics
    - Predictive model - showing strength of specific neighborhoods
    - Heat map visualization of neighborhood value strength
    - Report out of findings recommended applications

3) Walmart - Trip Type Classification

- Problem - Walmart wants to know why a shopper is coming into the store to enhance store sales

- Enabled decision - Stock items purchased together near each other enabling easy cross-selling
- Data available - https://www.kaggle.com/c/walmart-recruiting-trip-type-classification/data (Work4WalmarT)
- Approach
    - Using supplied trip purchase data, use market basket analysis to classify trips
    - Use trip types supplied by Walmart
    - Dimensionality Reduction (PCA)
- Deliverables
    - Report of models and visualizations used
    - Kaggle submission of predicted trip types

4) Prudential - Life insurance risk

Can you make buying life insurance easier?

In a one-click shopping world with on-demandeverything, the life insurance application process is antiquated. Customers provide extensive information to identify risk classification and eligibility, including scheduling medical exams, a process that takes an average of 30 days.

The result? People are turned off. That's why only 40% of U.S. households own individual life insurance. Prudential wants to make it quicker and less labor intensive for new and existing customers to get a quote while maintaining privacy boundaries.

By developing a predictive model that accurately classifies risk using a more automated approach, you can greatly impact public perception of the industry.

The results will help Prudential better understand the predictive power of the data points in the existing assessment, enabling us to significantly streamline the process.

https://www.kaggle.com/c/prudential-life-insurance-assessment/data