CSCI 379: Final Project proposal

What problem are you working on? Provide a paragraph background describing the problem.

Santander is challenging Kagglers to predict which products their existing customers will use in the next month based on their past behavior and that of similar customers. This will be used to improve Santander Group's product recommendation system.
Competition page is listed at
https://www.kaggle.com/c/santander-product-recommendation

What data are you going to use to solve the problem? Describe the data in a very general sense. Where did it come from? What does each object represent in the data?

The data is provided from Santander bank, through Kaggle platform. We are provided with 1.5 years of customers behavior data from Santander bank to predict what new products customers will purchase. The data starts at 2015-01-28 and has monthly records of products a customer has, such as "credit card", "savings account", etc. We need to predict what additional products a customer will get in the last month, 2016-06-28, in addition to what they already have at 2016-05-28. There is an important notice that this data does not contain any real Santander Spain customers, and thus, not representative of any actual geographical or societal demographics.

What are you going to do with this problem?

Since the data is especially large (approximately one million instances for the testing set alone), we don't rule out the possibility of using Python for this problem (especially when R processing gets slow). Python also has a variety of tools for statistics and numerical computing (with better efficiency) to deal with this problem and is itself a good learning opportunity.

- **Data cleaning:** Current data is not in ideal form and is in Spanish. We will have to translate attributes and information of the data into English so that it is readable. The trainting data set is also really large that we might need to explore sampling technique to reduce the size of the data and still have a close representation of the original one.
- **Visualization:** Explore the data to understand more about the demographics of Santander's customers. There are many different attributes and plenty of opportunities to visualize and bring insights about the problem.
- **Clustering:** Grouping similar customer groups using clustering technique.
- **Classification:** Use many different classification techniques to predict the outcome (additional products). We will explore a list of available methods first and hopefully be able to pick a large number of classification methods for this project.

How will you evaluate success?

Since this is a Kaggle challenge, we have a readily available evaluation method at Kaggle. We will use the Mean Average Precision @ 7 (MAP @ 7) method, which is described at the following hyperlink:
https://www.kaggle.com/c/santander-product-recommendation/details/evaluation
Since this is a classification problem, other evaluation methods already taught in class such as Confusion Matrix, ROC Curve can be used if suitable for the problem at hand.

One particular challenge for tackling this project might be the size of the dataset. 200MB of training datasets might be extremely challenge to run cross validation. We will find out different techniques to perhaps increase the validation speed and doesn't reduce the quality of evaluation process.

How will you get your work done? Give a reasonable list of milestones to reach, matching the stated deadlines above.

- **Milestone 1:** Explore the problem as well as the datasets of the problem. This will include data summary, data cleaning and some possible visualization.
- **Milestone 2:** Clustering customers into similar subsets
- **Milestone 3:** First predictions of products based on groups of customers
- **Milestone 4:** Devise evaluation metrics and methods to validate our product predictions
- **Milestone 5:** Improve the model based on our evaluation metrics