


Unit 7 - Final Capstone Proposal

August 08, 2018

Overview

For my final capstone at Thinkful's Data Science program I will be using a dataset from China's largest third-party mobile data platform, TalkingData. Understanding that everyday choices and behaviors draw a picture of who we are and what we value, we can use this information to optimize apps and manage marketing resources. The dataset contains information regarding app usage, geolocation and mobile device properties. My objective is to predict user's demographic characteristics based on the latter information, which in return gives us valuable information that will help millions of developers and brand advertisers around the world pursue data-driven marketing efforts which are relevant to their users and catered to their preferences.

Issues & Goals

1. **What is the problem you are attempting to solve?** TalkingData is China's leading third-party data intelligence solution provider: 80% of the Top 50 app developers in China utilize TalkingData to track their app metrics, analyze user data points, and optimize monetization. By predicting user's demographics we can manage resources and help developers to cater their apps according to their user's preferences. In other words, I'm attempting to save money to the company and its clients.
- 

2. **How is your solution valuable?** At the time the dataset was published, marketing resources (to name one aspect) were allocated randomly. Having the knowledge of demographic characteristics of each app user can save money and help to make wise decisions on where marketing efforts should be focused.
3. **What is your data source and how will you access it?** My data source is <https://www.kaggle.com/c/talkingdata-mobile-user-demographics> where a collection of files summing 1.3 GB is available for download, having 180,000+ unique user information and 110,000+ app properties. I will attempt to cluster according to both user's demographics and app information, thus evaluating which approach is more effective at the end.
4. **What techniques from the course do you anticipate using?** I will be pre-processing and cleaning the dataset using Pandas, balancing the data if needed, generating unsupervised features using perhaps TF-IDF and clustering with K-means models or other clustering methods. Moreover, I will be using feature selection methods such as SelectorKBest to optimize our training. As my specialization was aimed to Deep Learning, I'll run Neural Network models such as CNN (Convolutional Neural Network) or MLP (Multi Layer Perceptron) and evaluate their performance according to their accuracy score. However, if results are not the expected ones, I don't discard using supervised learning models such as Logistic regression, Lasso regression, Ridge regression or Support Vector Machines (SVM) to compare scores. Additionally I will use hyperparameters tuning techniques, such as GridSearchCV to optimize the model's performance.
5. **What do you anticipate to be the biggest challenge you'll face?** I think one of the biggest issues I will be facing is the processing power of my personal computer to handle this amount of data and run the mentioned models accordingly. If this results to be a significant issue I will probably migrate my workspace to a Virtual Machine at a cloud

service such as Google Cloud Platform having the advantage of using more computer power.

Conclusion

I believe the objective of this project to be of great value and could save time and funds to the company. However, I also believe that this information should be used as an additional tool for the respective departments that eventually could be involved, not being this the only resource to carry on decisions. I would suggest to deploy the results in a sample population of applications and test the results in the field for a limited time. If the results are successful, the company could deploy it in all of its applications.