Proposal

**1. What is the problem you are attempting to solve?**

I´m trying to enhance the knowledge that companies have of mobile phone users so that they can better target ads and marketing campaigns by segmenting population. This segmentation comes from predicting mobile phone user´s demographic characteristics by looking at the apps they use, the mobile phone and the apps they access and their usage in terms of time of the day in which they access, time, etc.

**2. How is your solution valuable?**

The solution is valuable as it will optimize (in the best case scenario) or at least provide with information that is relevant to the companies to make a better use of their resources when they are addressing marketing campaigns. If the quality of the leads that companies are getting thanks for this information is higher by better allocating resources that will bring a reduction and better use of economic resources. It will not only reduce the cost of the ads and marketing campaigns but it will also increase revenues increasing the overall margin and conversion.

My assumption is that the demographic features that are relevant for marketing campaigns will change by regions but this data will allow me to build the first model and then test and tune across regions.

**3. What is your data source and how will you access it?**

The data sources that will be used are:

1. An available dataset from TalkingData in Kaggle that targets 70% of the Chinese market. [https://www.kaggle.com/c/talkingdata-mobile-user-demographics](https://www.kaggle.com/c/talkingdata-mobile-user-demographics" \t "_blank)
2. Data regarding basic demographics in China and data regarding data consumption from the UNSD Demographic and social statistics. <https://unstats.un.org/unsd/demographic-social/index.cshtml>

**4. What techniques from the course do you anticipate using?**

The process that will be followed for this capstone is:

* **Data Wrangling and cleansing**: Removing blanks and NaN and running boxplots to identify outliers. Outliers will be dealt with, at this stage removing the ones that are distorting the dataset.
* **Exploratory Data Analysis:** Identify if there is class imbalance by running some basic statistics. Visualizations will be done to understand the relationship between the variables (if any) and the behavior of each of them. Correlation between data will be examined.
* **Data pre-processing:** Feature generation and reduction will be done based on PCA to understand the maximum number of features and then RFE, feature importance using random forest and feature selection using KBest to see which features are the most relevant and if the results are consistent between the different methods. Furthermore, based on the results of the correlation analysis and behavior of the data, new features will be generated and tested to see if they are more or less relevant following the procedure previously outlined, Correlation will be tested again and visualizations provided to understand the features (relationship between the set of selected features, explanatory power, correlation matrix, heatmap). Lasso and Ridge regressions will be used to test the final set of features and their relative importance. By the end of the data cleansing stage, features will be clean, normalized and ready to be used in different models. Shifting gears into the outcome variable, classes will be balanced through resampling techniques, either down sampling the majority of up sampling the minority. The balancing of each class will come from the exploratory data analysis that will be done when cleansing the data. This will avoid overfitting when the algorithms are run on the test data. The pre- and post- resampling size of classes will be shown through different visualizations showing the behavior of each of them. Data will be split 75/25 between the training and the test set.
* **As part of the second part of the Exploratory Data Analysis** (after the data has been cleansed), clustering will be done using kmeans (providing the algorithm with the best number of clusters), mean shift, spectral clustering and affinity. All of them will be compared against each other to determine the one that best represents data.
* **Models & Validation:** All models will be run in the training set and validated in the test set. In the training set, hyperparameters of the models will be used using gridsearch setting the score to the overall accuracy and cross validation with five folds. The time to fit the models will be measured to see their adequacy for production. A classification report, confusion matrix, the overall accuracy and % of type I and II errors will be measured in the test set for validation. The initial list of models that will be run are:
  + **Logistic Regression:** C parameter will be tuned to obtain the best accuracy. The penality used will be L1 or L2.
  + **Naïve - Bayes (multiclass**) for multiclass classification
  + **KNeighbors Classifier**: tuning the number of neighbors against the overall accuracy of the model and weights used.
  + **Random Forest**: For this model the parameters tuned will be N\_estimators determining the number of trees that will be part of the algorithm. Max depth determining the size of the tree.
  + **Decision Tree**: This model will run to compare the results obtained through it with the results obtained using Random Forest. In this max leaf nodes will be tuned accordingly.
  + **Support Vector Machine**: Parameter C and type of kernel will be tuned.
  + **Gradient Boosting Classifier**: As it usually is less intensive than random forest, it will be run. The parameters that will be tweaked are the number of estimators, giving the number of weak trees that the ensemble model would be using the maximum depth of each tree giving the number of nodes in each of the trees.
  + **Neural networks** (MLP, Convolutional and recurrent networks). This model will have a double use. On one side it will be used to generate features and on the other side for the supervised problem of multiclassification. In the second case, the number of neuros, hidden layers size, activation functions and dropouts to fight overfitting will be tuned.

All models will be validated on the test set 25% of the data as specified at the beginning of this section,

* By the end of the pipeline, visualizations will be provided with the insights obtained from the models and the best model will beselected based on the overall accuracy of the model. All of them will be compared to the results obtained using PCA but as we need the features to describe the characteristics of the potential clients and to cluster them, PCA will not be used as features for the model. Clusters based on the features will be run using the abovementioned techniques and visualizations will be provided.

**5. What do you anticipate to be the biggest challenge you’ll face?**

The three main challenges that I foresee are:

* Finding the best set of features so that explanatory power is maximized will be the main challenge. If explanatory power of the features found is not stable then RFE will be run for each algorithm to see which features should be eliminated as they can be adding noise. Once compared to the results obtained using PCA, the distance in accuracy points between them will be used to test the features that have been generated.
* The selection of the best model from a production and accuracy standpoint. Although the most computational intensive algorithms can provide the best accuracy, it can be difficult to scale them up to run the process in real time.
* Computation power: This is a restriction that I have found through all the data science bootcamp so I´m expecting to reset the virtual machine and run it in cloud. In some cases, the model could take up to 10 hours running it in google cloud with a v8CPU and 30 GB.