**Project Proposal**

**Group name: Do Not Overfit**

Andy Huang (yh3090)

Xiaoxi Zhao

Yiming Tan

**Main question**

* Build a predictive model to predict those policies that are most likely been canceled during the effective term in insurance industry
* Find out the key drivers that cause policy cancellations

**Reasons for choosing this topic**

* This topic is a competition held jointly by UConn Statistics Department and NESS NextGen, we want to join the competition and to see if we can get award
* This topic is a widely discussed topic in insurance industry as well as in any fields related to customer relationship. By doing this project, we can improve our modelling skills as well as our business acumen

**Data Type**

* The data we use is the 4 years of property insurance policies from 2013 to 2017, which are provided by the organizer.
* If it’s needed, we will hunt for extra data online.

**Work Plan**

* Programming Language: Python
* Target Problem: Binary Classification (imbalanced)
* Metric Used: AUC
* Data preparation: Encode categorical data into numerical value (one hot encoding, label encoding, mean encoding etc.) to fit model. Use PCA for dimension reduction if necessary.
* Interpret the results with rather simple models ( instead of stacking), for example: logistic regression, SVM, random forest, gradient boost decision tree, Adaboost, Xgboost (May use L1, L2 loss in linear models and variables selection techniques if necessary)
* Increase test AUC as high as possible by using stacking methods or deep neural networks
* Read relevant papers to get some intuition about features engineering, feature selection and model improvement
* Create a poster to present our outcomes