**SEIS 763 Machine Learning**

**Project Proposal**

**Due: 04/10/2019**

**Submitted by**

**Ke Chen**

**Jun Li**

**Jared Oeth**

**Satish Dandayudhapani**

**SEIS 763 Project Proposal PetFinder.my Adoption Prediction**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

For our class project, we intend to use the Kaggle competition Pet Finder dataset. Millions of stray animals suffer on the streets or are euthanized in shelters every day around the world. If homes can be found for them, many precious lives can be saved – and more happy families created.

PetFinder.my has been Malaysia’s leading animal welfare platform since 2008, with a database of more than 150,000 animals. PetFinder collaborates closely with animal lovers, media, corporations, and global organizations to improve animal welfare. Animal adoption rates are strongly correlated to the metadata associated with their online profiles, such as descriptive text and photo characteristics. This dataset will help answer the questions which will guide shelters and rescuers around the world on improving their pet profiles appeal, reducing animal suffering and euthanization.

The dataset is about of 2.7gb of size and has 24 variables with a mix of continuous and categorical. Main dataset contains all important information about pets: age, breed, color, some characteristics and other things. The dataset also includes images and sentiment for each pet. There are few features with missing values in the dataset i.e. images, sentiment, this will be addressed appropriately during project execution. This is a classification problem and the target variable of the dataset is Adoption speed. The value is determined by how quickly, if at all, a pet is adopted, and it ranges from 0-4 where 0 being the fastest and 4 being the slowest adoption rate. The detail about the dataset is shown in the link below:

<https://www.kaggle.com/c/petfinder-adoption-prediction/data>

Jared and his family are in the process of adopting a pet and he brought the idea of using this dataset for the class project which interested everybody else on the team. Below are the few questions which we are focusing on to explore, analyze, visualize and predict the target variable.

* Data exploration - Exploring the features and their interactions – specifically, what are some of the most important factors impact the adoption rate.
* Comparing distribution of features in train and test data
* Developing algorithms to predict the adoptability of pets – specifically, how quickly is pet adopted?
* Trying various types of feature engineering
* Trying various models and comparing the accuracy of each model on test data

We will be using Anaconda Python distribution for preprocessing and data exploration tasks and Google Cloud Platform (GCP) for model training and testing. Some of the methods we may use for the project include EDA (Exploratory Data Analysis), LGBM, XGBoost, and Ensembling of different models. The deliverables include a project report about 10-20 pages plus code and will be submitted electronically on canvas.