Big data project proposal

Team 15

Name	Sec	B.N.
Reem Emad	1	33
Osama Magdy	1	14
Ziad Atef	1	35
Yousef Gamal	2	39

a. Idea:

Millions of stray animals suffer on the streets or 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. Animal adoption rates are strongly correlated to the metadata associated with PetFinder.my online profiles and dataset. This problem aims to predict the speed at which a pet is adopted.

b. Dataset(s):

PetFinder.my Adoption Prediction | Kaggle

c. Planned approach or Proposed solution:

- First we may use PCA in order to focus only on the most important features.
- There is a class imbalance. So, a method to deal with class imbalance (smote or cost-sensitive learning for example) will be used.
- We will use different features as feature vector to train Naive Bayes and Logistic regression classifiers (maybe changed later)

→ For Language, We are going to use Python with the Spark framework.
→ Map Reduce in ML pipeline parallelization in naive bayes.