Insurance Risk Analysis

Harsh Pandey, Mark Tiburu, Sarang Wardadkar (Team 8)

Data

- Kaggle Competition
- Prudential Life Insurance Assessment
- https://www.kaggle.com/c/prudential-life-insurance-assessment

Goals

1. Develop a predictive model that accurately classifies risk

2. Apply and compare results for various Machine Learning Algorithms

3. Use Scala to stream and clean the data set

4. Getting acquainted with MLlib

Data Clean Up

- Missing Data
- Dimensional Reduction using Filter methods in Scala
- Formatting and validating

Linear Regression

XGBoost

Decision Tree

What We Aim to Achieve

- To successfully predict risk for customers given various inputs

- Implement machine learning algorithms

- Learn to use Spark in Scala

Repository Link:https://github.com/swardadkar/CSYE7200-Fall2017

Time Lines

- Data Clean Up and In Depth Understanding: 1.5 weeks

- Implementing Algorithms: 1 week/each

- Clean Up and Final Presentation: 1-2 days

Stretch Goal

- Implement Logistic Regression

- User Input based streaming (Form)