Compensataion Prediction

Team #13

Junyuan Bao
Boyu Chen
Yu Fang

Goals

When people devastated by a serious car accident, we can use this model to help users' family, friends and other relationships to predict the mental and time emergency spent.Help them improve their claims service for households among the victim.

After testing our models through the validation dataset, we want to achieve a good model to use our model help users to accurately predict the damage they suffered and compensation they can claim.

Use Case

User inputs queries (damage range, accident type ...) and receives lists of matching records;

System fetch train and test datasets from sources and build models for future analysis. After set up the model, system runs data on the learning machine model and displays list of feature vectors, each with predict labels.

Users click on the request button and type in details into the system which will evaluate their time and mental emergency spent.

Methodology

Using Gradient Boost Tree, Random Forests, neural network or other supervised machine learning methods.

Using Spark MLlib decentralized machine learning framework.

Data Source

RAW Data Set (for the data privacy, Allstate denied to provide the specific information, only shown by category or continuous data type plus data id):

https://www.kaggle.com/c/allstate-claims-severity/data

Data quantity: 188k rows of training dataset

Milestones

Week 1	Do research on Spark and construct a skelonton of project
Week 2	Implement crawler features and clean data
Week 3-4	Implement machine learning features and define a sensitive result set
Week 5	Do test case and validation, add documentation

Time will be adjusted accordding to concret task volume.

Program Part

Data ingestion: ingest data from original kaggle dataset and reference claims severity category dataset.

EDA: apply a set of rules to each data entry, reject those entries that do not qualify. The system also need to build a set of aggregate measurements for the dataset and integrate the data set from distinct sources as a whole.

Data analysis: build two distinct data model based on "random forest regression" and "gradient tree regression" models. Our system will also perform some analysis to compare the accuracy result got from the model.

Program Part

Code repository (link below):

https://github.com/FOURTHZI5/CSYE7200_Scala_Final_AllState_Severity.git

Acceptance Criteria

After the project, we will get a loss value prediction based on the training dataset and be evaluated on the mean absolute error against actual loss.

Reach a "Mean Absolute Error(MAE)" over 1100. (Measured by Kaggle)

Our system will also show the performance gap between the "random forest regression" and "gradient tree regression" models.