**Allstate Claims Severity**

“Allstate is currently developing automated methods of predicting the cost, and hence severity, of claims. In this recruitment challenge, Kagglers are invited to show off their creativity and flex their technical chops by creating an algorithm which accurately predicts claims severity. Aspiring competitors will demonstrate insight into better ways to predict claims severity for the chance to be part of Allstate’s efforts to ensure a worry-free customer experience.”

Link: <https://www.kaggle.com/c/allstate-claims-severity>

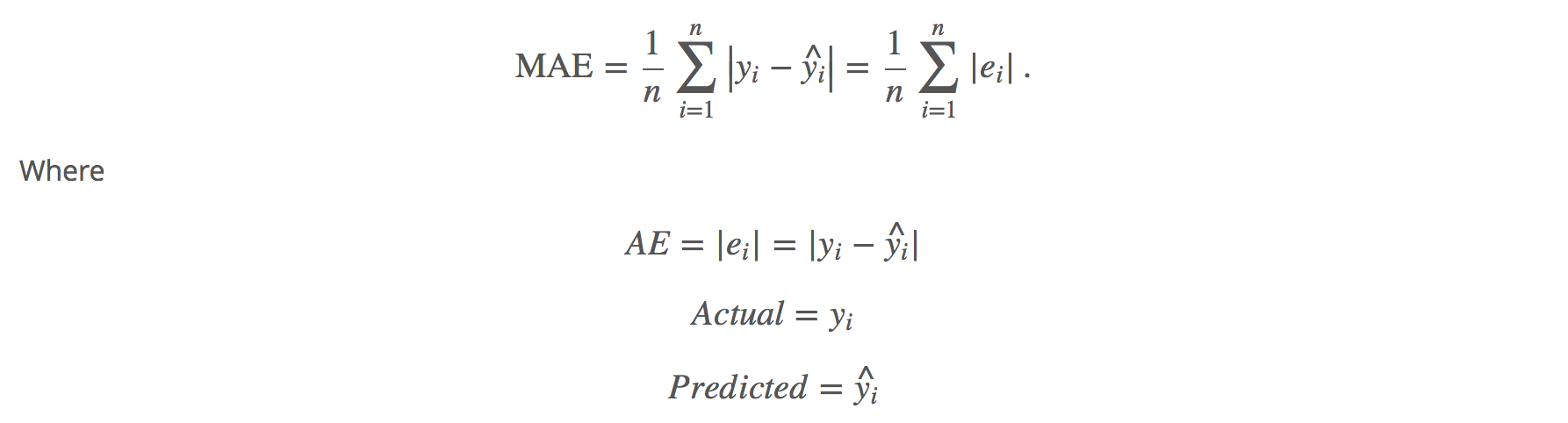
**Introduction**

This is a **live** Kaggle Recruiting competition that started on October 10. It will finish on December 12, so there will be a chance for you to see how well you did in the near future! Just like with Facebook, Allstate has hosted on Kaggle in the past and they are back with another ML challenge.

In the train.csv, you will find 131 columns of attributes, mostly categorical data (and the column labels conveniently tell you which data type they are), along with a response column called “loss”. You are directed to predict the amount of potential loss given a claim according to the features that they provide. Importantly, Allstate does not provide any information about what each feature represents, so this competition lends itself well to just practicing ML techniques.

**Objective**

The metric for this competition is the Mean Absolute Error:



Description: <https://www.kaggle.com/wiki/MeanAbsoluteError>

This is very similar to the RSS (Residual Sum of Squares) except we are taking the absolute value of the difference in predicted vs actual y values rather than the square. We are also averaging this difference across all rows instead of summing.

As before, in addition to the accuracy, you and your teammates will be drafting an ipython notebook detailing your data cleaning, exploratory analysis, and modeling methods.

**Logistics**

Project submissions (notebook and Kaggle score board submission) will be due by **Wednesday 11/30 11:59pm.** Please make sure to record your public leaderboard score somewhere noticeable in your notebook and please submit them on to Gradescope.

**Notebook Guidelines**

1. Follow the general structure of notebooks you see on Kaggle:
   1. Basic data exploration with graphs, summary statistics, and outlier detections
   2. Data cleaning/**Feature Engineering\***
      1. A note on feature engineering: Since we are not provided the column labels, we will be very lenient with the grading of this proportion. However, to push you all in an interesting direction, it might be helpful to read through this notebook to get some ideas (EDA stands for Exploratory Data Analysis):
      2. <https://www.kaggle.com/dmi3kno/allstate-claims-severity/all-the-allstate-states-eda/comments>
   3. Model building
   4. Model evaluation
   5. Model interpretation
2. If you are in part of a team, you can split the work per section – but the best way to learn is to help do a little of everything. There are a lot of features and no person will produce the exact same analysis.
   1. Tip: This notebook is **absolute gold**:
      1. <https://www.kaggle.com/sharmasanthosh/allstate-claims-severity/exploratory-study-on-ml-algorithms>
   2. It contains, essentially, the validation scripts for the majority of commonly used ML algorithms plus more! Wish I had something like this when I first started Kaggling…

**Grading**

We will set our own benchmark accuracy score and your model **MUST beat it to pass the class.** ☺

Points are distributed uniformly per team as follows:

70 points – Benchmark = (1217.52141 Public Leaderboard Score)

.

.

100 points – Top team model.

Don’t actually have the exact script for the above benchmark since this is the one Kaggle internally generated. BUT, the model used is Random Forest so I would imagine something similar to Santhosh’s script (in the tip section link) produced that score.

**The notebook will be graded for effort and quality** on a 0-50 scale (10 points X 5 sections in the notebook structure).

For each section, point values are outlined below:

1. Data Exploration

2 points for graphs of any kind

2 points for statistics of any kind

3 points for thoroughness (did you analyze enough to get a good understanding of the data?)

3 points for quality/creativity

1. Data Cleaning & Feature Engineering

5 points for cleaning the data of the most apparent problems (e.g. missing values, outliers, dummy variables, etc.)

2 points for feature engineering any variable (**Just show that you tried to do something. It does not have to be in your final model)**

3 points for “creative/unique” feature engineering

1. Model building

5 points for setting up the train and test sets correctly

2 points for successfully running a model

3 points for successfully running more than one model

1. Model evaluation

7 points for correct out-of-sample test accuracy to the public leaderboard accuracy (within a reasonable threshold)

3 points for demonstrating that your local test accuracy is correlated with the public leaderboard accuracy

1. Model interpretation

4 points for interpretation of any kind

3 points for thoroughness (did you fully grasp what your model is telling you about the data?)

3 points for comparing different models