Team: 06 Codechella

Members: Sofia Blavatsky, Joshua Pardhe, Noah Ruiz

Executive Summary

In 2014 researchers from Yale, with collaboration from play2PREVENT developed the “Elm City Stories” interactive role-playing game, with the intention of educating youth about the dangers of risky behaviors such as STD’s, unprotected sex, underage drug consumption and more. Flash forward to 2022 and with the increase of computational power, there has been increased interest from the research team regarding actual game log data. Because of this in collaboration with the American Statical Association, researchers from the “Elm City Stories” study have given teams such as ours access to the game log data.

To narrow the projects scope and purpose, our team decided to focus on trying to see if we can predict player skill based on overall interaction. In terms of researcher interest, being able to understand player skill based on how engaged they were in the game can be a very powerful tool, for future studies and analysis. Before analyzing the data, our team, used the three sample player files and to try and identify any trends within each respective sample players game log data. What our group found was quite interesting, when analyzing behavior at a micro level there was a clear correlation for certain players between time elapsed playing the game, and actual skill point increases. Additionally, when our group analyzed, time elapsed playing the game at a macro level we found that the correlation coefficients dropped significantly, which can be attributed to the random nature of time elapsed in the game. Further, when graphed at a macro level there are clear clusters between players, depending on how much time they spent on the game, with players that spent around 9.079 hours being grouped in a hockey stick like cluster and all other players being in various clusters far from the median ‘time elapsed’.

To do this our team needed to decide how we would classify ‘engagement’. Our team decided to define engagement mainly in terms of total sums and counts of ‘event_id’s’, ‘animatic_time elapsed’ and ‘sense_id’s’. Boiling those variables down, our team decided to use various ‘event_id’ as our predictor variables because, each event_id within our model itself corresponds to a game action. Event_id 210 and event_id211, were of particular interest, mostly because each id corresponded with a player engaging with the game using their “Senses” and various other skills points. Sense Id’s were chosen as a form of engagement because each sense_id essentially indicates when an object is clicked or in terms of the data dictionary, “Selecting a sense makes different objects activated when clicked”. In a way our team assumed that using counts of sense and event id’s could be used as a form of measuring user clicks and choices. Additionally animatic time elapsed was chosen as a measurement of engagement because it tracks the total amount of time elapsed in each animation, which our team hypothesized could lead to an increase in player skill. Regarding player skill, our team decided to create an averaged-out skill column called ‘Accumulated Skills’, which is simply the sum of all skill categories dived by the count of skill categories. Running correlation test, our team found that ‘event_id210’, ‘event_id211’, ‘sense_id0’ and ‘animatic_time_elapsed’ all had a high positive correlation ($r>0.60$) with ‘Accumulated Skills’. Finally, our team built out a multivariate regression model, that used the values with the highest correlation coefficients as parameters, which reported an adjusted R-square of .50 and statistical significance for all model parameters at an alpha of 0.05. However, this suggests further research in the actual model parameters itself as the model has poor utility.