CSE 546 Milestone

Langley DeWitt, Zoheb Siddiqui

Data Preprocessing

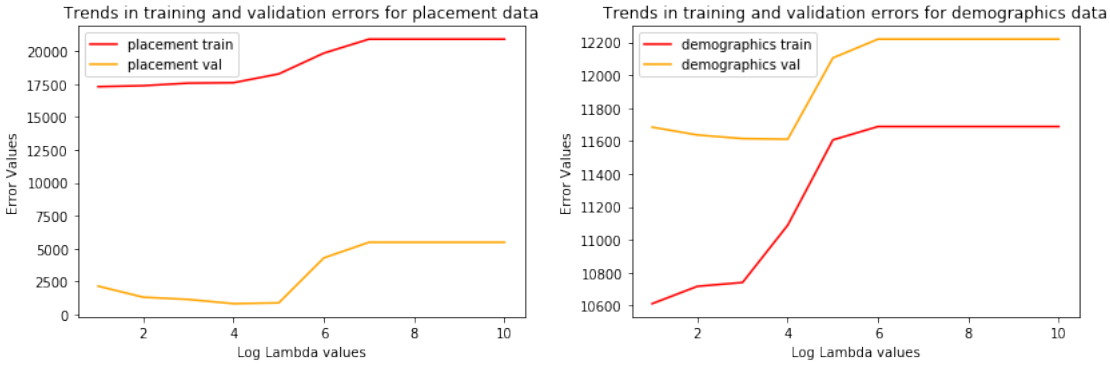
Most of the features used were categorical and since we are performing a regression task, it is necessary to one-hot encode the data.
All categorical data except for 'age range' were one-hot encoded. Age range was split into 'lower age' and 'upper age.'
Then the data was split into train, validation and test sets (70%, 15%, 15%). We decided to perform this split instead of doing K-fold Cross-Validation as we are dealing with time series data.

Models

After preprocessing the data, our next task is to figure out which regression model we shall use to accomplish our goal. We decided to test lasso regression, ridge regression and elastic net regression. KNN and Kernel regression techniques were avoided as we are dealing with time series data.
For each model type, we created 1 model for placement data and one for demographics data. The best hyper parameters were found by checking RMSE on the validation set.

The results from the test are below:

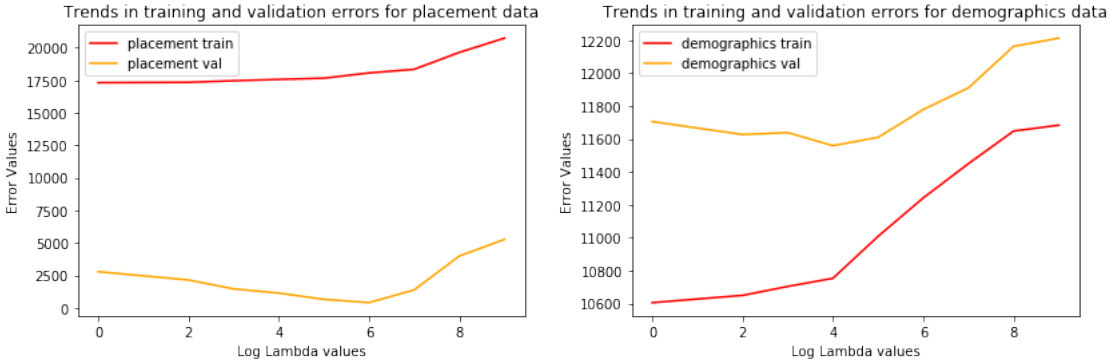
For Lasso regression, we tested $\lambda = 10^i \quad \forall i \in [1, \ldots, 10]$. The best results for both the placement and demographics models were achieved for $\lambda = 10^4$



Best placement lasso RMSE: 804.4208754557318
Best demographics lasso RMSE: 11611.273957138816

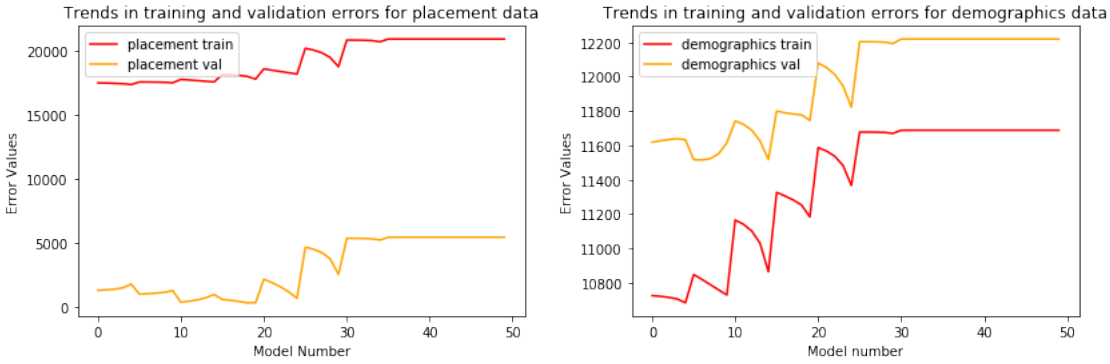
For Ridge regression, we tested $\lambda = 10^i \quad \forall i \in [1, \ldots, 9]$. The best result for the placement model was achieved for $\lambda = 10^6$. The best result for the demographics model was achieved for $\lambda = 10^4$.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.



Best placement ridge RMSE: 408.7801324475013.
Best demographics ridge RMSE: 11559.54490341586.

For Elastic Net regression, we tested $\lambda = 10^i \quad \forall i \in [0, \ldots, 9]$ and $L_1 \ ratio = [0.1, 0.3, 0.5, 0.7, 0.9]$. The best result for the placement model was achieved for $\lambda = 10^3$ and $L_1 \ ratio = 0.7$. The best result for the demographics model was achieved for $\lambda = 10^1$ and $L_1 \ ratio = 0.3$.



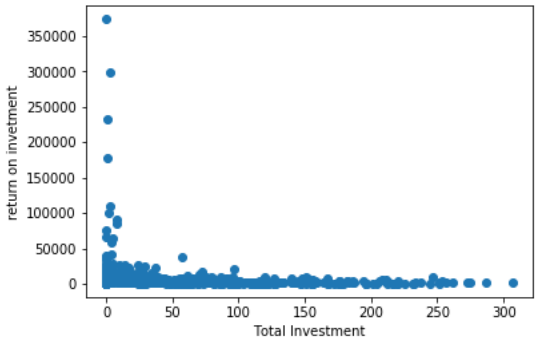
Best placement elastic net RMSE: 356.8582278583902
Best demographics elastic net RMSE: 11514.83827710379

Based on this information we can conclude that the above identified Elastic Net models fit the data best and these models will be used for all future predictions.

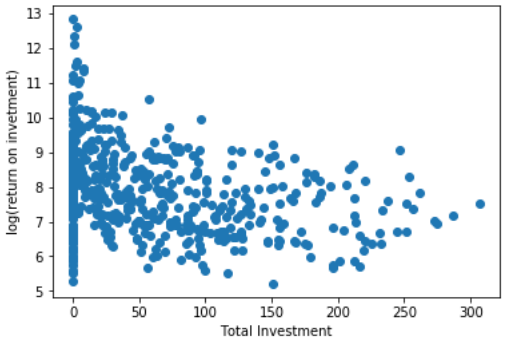
Limiting

We expect that as we invest more heavily into certain demographic groups or into advertising in certain places, we will get diminishing returns. This is because if the group is small, everyone in that group will see the ads and advertising more to the same people will be less effective. this means that we need to account for that in our model if we want to predict the return on investment. To do this we are adding a new feature that keeps track of how much has been invested into a group. The way the feature is calculated is that all of the investment into advertising to the group for each ad campaign up to the date in question is summed up. When this feature is graphed against return on

investment for a single day we get the following graph.



This graph show that higher returns on investment are more likely when the total investment is lower. This supports the idea that investment in the pas reduces the return on further investment. When we look at the log of the return on investment against the total investment we get the following graph.



This graph seems to show a linear relationship, so $exp(total\_investment)$ may be a useful feature in linear regression.
What we have left to do here is to use this to find demographic specific drop offs for return on investment. This shows that on average return on investment for a demographic will decrease with total investment into that demographic, but we would like to know how much a specific demographic's return on investment will decrease. To solve this we will likely introduce more features that will encode this. For example, a feature that holds the total investment into advertising to men for data points representing men and zero for women. If this is implemented for more all demographic categories, it could give us the behaviour that we want.

Appendix

The code for data processing and model testing can be found **here**