CSE 546 Projet

Langley DeWitt, Zoheb Siddiqui

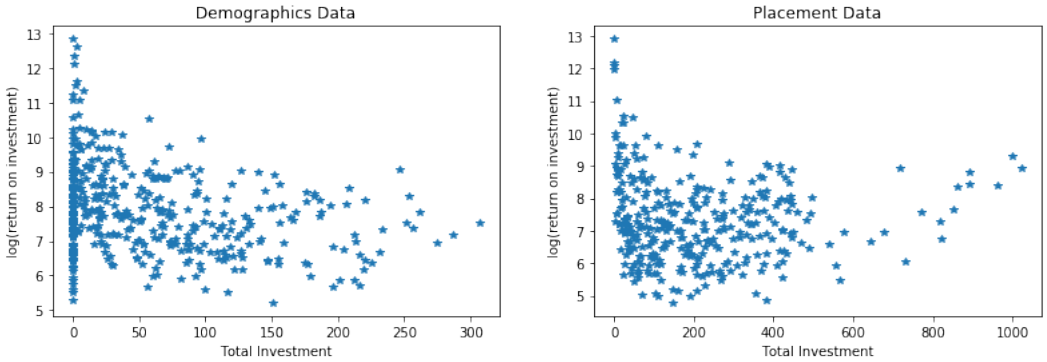
Introduction

The purpose of this project is to create a machine learning model that predicts ad revenue. We have two data sets corresponding to the same set of advertisements. We shall first perform some feature engineering and validate those features.

Once this is done, we fit the chosen data set to various regression models and perform validation to determine the best model and hyper-parameters. We then test the best models against the test data to measure their accuracy.

Feature Engineering

The first step we took to build on the models that we already had was to introduce a feature that could help prevent us from to heavily investing in advertising to a single, limited audience. We wanted our model to be sensitive to how much we had already invested into that audience, so we created a feature to track total investment into a demographic or placement. To do this, we added up the amount spent feature(which is the amount spent on advertising to the group in a day) from all the previous days of the add campaign for each combination of categorical identifiers that we had in the data. If we graph the feature against the log of the return on investment for both the demographics data and the placement data, we get the following plots.



There appears to be a trend in the demographics data where larger total investments seem to correlate with small returns for that day. In placement data its less clear. To do a more quantitative analysis of this feature, we did a simple linear fit to the data, for all of our features, on the two data sets with and without the feature, and got following errors on the validation set.

Root Mean Squared Validation Error		
Data Set	With Total Investment Feature	With Total Investment Feature
Demographics	10605.07	10611.8
Placement	17310.4	17831

This appears to show that this feature does give a smaller error. In the placement data case, the gain in accuracy is larger, but there is a slight improvement in both cases. This tells us that this feature is more likely to help our model than to hurt it.

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

Saturation Fitting

We expect that as we invest more heavily into advertising to a specific demographic or in a specific place, we will get diminishing returns on our investment. This is because if a person has already seen and responded to an ad they are unlikely to do so again. The way that we modeled this was by assuming that when we invested money into advertising to a group some fraction of them would see it and be less interested in the future. This results in the following model.

$$return = e^{w(p)total\_investment+b(p)} \tag{1}$$

In this model, the values $w(p)$ and $b(p)$ are the rate which the return on investment decreases with the amount that a group has been advertised to and the return on investment that we would get at the very beginning, before the group as been advertised to respectively. Both of these are functions of p the population. For different demographics and placements, these values will be different. This is what we hope to learn with our model.
The first thing we did here was to start looking at the log of the return.

$$log(return) = w(p)total\_investment + b(p) \tag{2}$$

This way we can use linear regression techniques to find possible functions for w and b. However, this did have the downside of forcing us to not use a lot of our data. Because many of the returns on investments were 0 we could not fit to their log. As a result, we were unable to use this as a complete model for our data, but instead as a way to correct for the effect of saturation and then fit another model on all of the data. The end result would be a predictor of the following form, where $m(x)$ is the model that we fit on the corrected data using both demographics or placement data and data on how much money is given to an add that day.

$$return = m(x)e^{w(p)total\_investment+b(p)} \tag{3}$$

To fit to fit the $w(p)$ and $b(p)$ we converted the features to one-hot encoding, and then used regularized linear regression to fit the data. To regularize the the norms of the w and b vectors we used different hyper parameters. This is because they are two distinct parts of the model that could be more or less sensitive to the data that we put in. For example the following formula would be objective function for ridge regression where T is a diagonal matrix with the total investment feature as the diagonal and p is a matrix of one hot encoded data points.

$$w, b = argmin\|Tpw - log(y)\|_2^2 + \lambda_1\|w\|_2^2 + \lambda_1\|w\|_2^2 \tag{4}$$

Mean Squared Validation Error

Model Type	Demographics Data	Placement Data
Constant	1.56	1.93
Linear Regression	1.60	8.71
Ridge Regression	1.45	1.77
Elastic Net	1.37	1.85
Lasso Regression	1.30	1.85
lasso filtered Ridge	1.48	1.85
lasso filtered Elastic Net	1.38	1.85

The table shows that in each case the lasso model works well relative to the other methods, and the models that were fitted after features were removed using the lasso model have about the same error as when they were fit with all the features. This implies that for a linear model only some subset of the features are actually helpful in fitting the function.

In addition it shows that the the models were, in general, not doing that much better than using a constant function. This can also be seen in the hyper-parameters that gave the best validation errors. The regularization factor for the w vector tended to be quite large (between 500 and 1000), indicating that models with low dependence on the total investment feature were favored. This is an issue because that is the behaviour that we wanted to model.

Finally we tried to use this model to correct for decline of the return on investment, and found that it did not help. The way that we tested this was we used the model with the lowest validation error(in both cases the lasso model) and divided the return on investment value with the

output of the model. The idea being that if the model fit the phenomenon well this would remove the effect from our data and improve the accuracy of our models that we fit on that data. The error was the same or worse with when we did this than when we fit to the unchanged data. This tells us that the relationship between w and b and the population is not linear in each of the categories. In other words, there is more interplay between the features, than our model accounts for. The impact of being in one category can be influenced by what other categories the population falls under.

Final Model

Having concluded that the engineered *total investment* feature provides a more accurate fit and that the *saturation fit* didn't help improve our model we decide to include the former and not the latter in our data sets.

With our finalized data set we create various regression models and use a validation sets to find the best hyper parameters. We compare the best models of each regression type to determine which regression models perform best on both data sets.

Once this is done we measure the accuracy of the best models against the testing sets and report it.

We tried Lasso Regression, Ridge Regression and Elastic-Net Regression to fit our data sets. The results from the fits are given below.

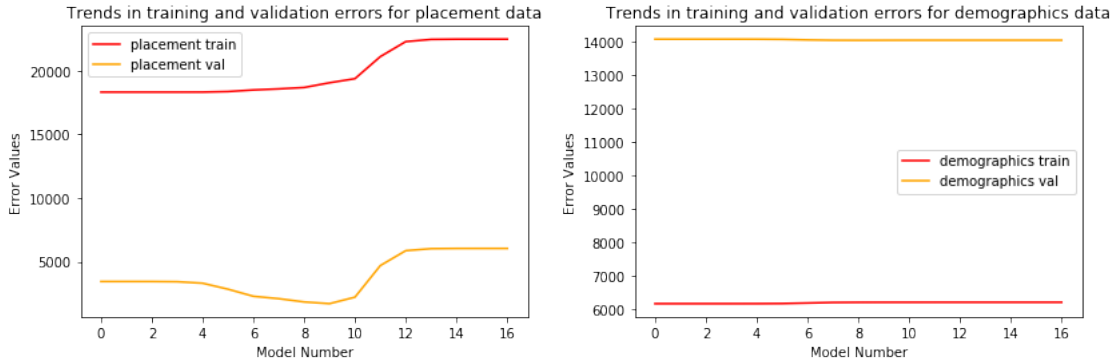
Lasso Regression

For the placement data set the best λ value was: 100000, and the corresponding validation RMSE was: 1793.33.
For the demographics data set the best λ value was: 10000, and the corresponding validation RMSE was: 14042.20.

Ridge Regression

For the placement data set the best λ value was: 1000000, and the corresponding validation RMSE was: 1710.08.
For the demographics data set the best λ value was: 1000000, and the corresponding validation

RMSE was: 14039.37.



Elastic Net Regression

For the placement data set the best λ value was: 1000, the best l1 ratio was: 0.7, and the corresponding validation RMSE was: 1705.16.
For the demographics data set the best λ value was: 100, the best l1 ratio was: 0.7, and the corresponding validation RMSE was: 14039.66.

Thus, we decide to use Elastic Net model for the placement data and Ridge for the testing data. The testing RMSE for these models were 11996.44 and 4064.94 respectively.

Combining both models

Since our data is fragmented we had to create two separate models to predict the revenue however a complication in this process is that both data sets share some features and thus, simply adding up the RMSE for both models is not an accurate measurement of the RMSE.

We decided to create a third model with just the features that both data sets have in common. We calculate the testing RMSE of this model after the best one has been identified through validation. The best model identified through validation was an Elastic Net model with the best λ value as: 0.001, best l1 ratio as: 0.5, and validation RMSE as: 1195.36.

We then use the inclusion-exclusion theory to calculate the true value of the testing RMSE. According to our hypothesis the RMSE that arises due to the placement and demographics models includes the RMSE from the third model twice. Thus, subtracting this RMSE from the sum of the first two should give us the true RMSE. The test RMSE for the third model was 8709.21.

The testing RMSE for the combined model was: 11996.44 + 4064.94 - 8709.21 = 7352.17

Appendix

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