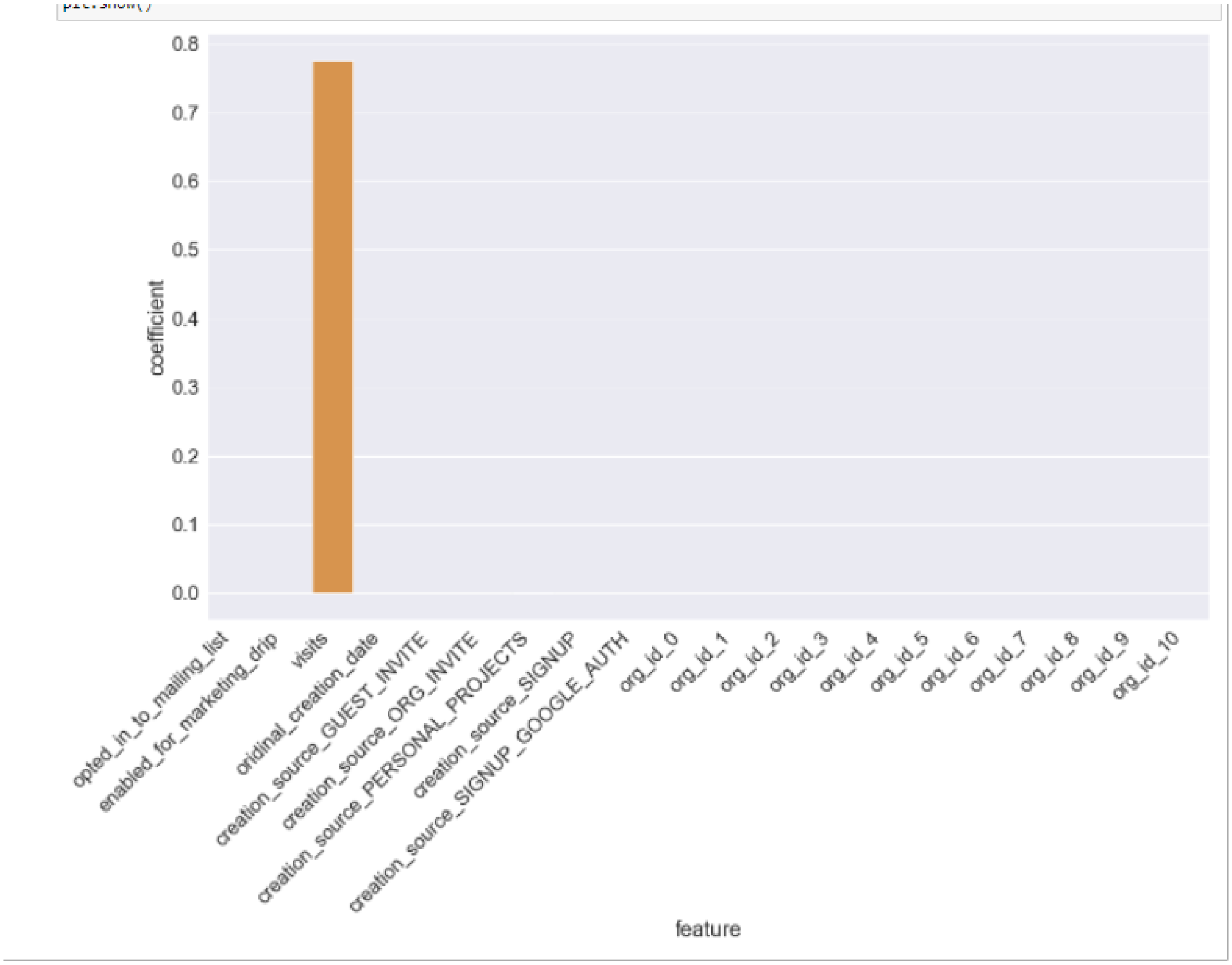
What did we do?

This project was mostly an exercise in feature engineering. I created two features, which were the the ‘adopted’ and ‘visits’ column. Adopted was a our target variable, simply a boolean value, true if they’ve visited three times within a week period, false if not. I also tacked the number of visits they had total onto the main dataframe.

For pre-processing, I did three notable things. 1. I encoded the account creation date as an ordinal, figured it was fine for the model to treat time as a continuous variable.2. I dropped columns that were going to be unique per user account, like email, name and last_session creation_time 3. I one hot encoded categorical variables like creation_soruce, org_gid and invited_by_user_id.

What did we learn?



```
plt.show()
```

I ran a logistic regression model on my new dataframe, and then extracted the feature importance. With these features, the only important feature was # of visits. It accounted for almost 80% of the variation, the remaining 20% being distributed across more than 2000 Features.

Things I’d like to do, if given more time:

- Add more features: I would have loved to add features like # of visits before adoption, or avg time between visits
- Make a model without the visits column and explore it’s accuracy
- Deeper EDA
- Grid Search and Cross Validate the hyper parameters, as it was I just went with the defaults as they’re going to be effective enough for preliminary EDA. A prod model I would prefer to have done as close to an exhaustive search of parameters as possible.