Predicting Engagement by Category

General Assembly Part Time
Data Science Final Project

Alex Rees

AGENDA

Problem Statement
What is the project objective?

Data Sources
When, where, and how was the data collected?

Feature Engineering
Creating the features to us in modeling and exploratory data analysis

Modeling
Finding the model that works best in solving for the problem statement

Inference
Final thoughts on the results

PROBLEM STATEMENT

Launch Forth is a co-creation platform on which companies launch rapid product design challenges. Launch Forth's community of 200k+ engineers and designers submit product designs and compete to win the challenge and have their design manufactured. Engagement rates within these projects are a primary selling point for new business. This project's objective is to explore if can we predict engagement rates based on a project's category?

Hypothesis:

Engagement can be predicted by category, and projects in the Ground Mobility category will have the most engagement.

DATA SOURCES

All 5 datasets were pulled from the Launch Forth platform and represent a sample of all actions taken on the platform since October of 2016.

Projects
List of all projects and respective categories

Watches
A 'follow' on any content other than a project

Follows
Follows specific to a project

Posts
A comment, which can be made on any piece of content or project.

Entries
Posted to challenges

Ideas
Posted to brainstorms

FEATURE ENGINEERING & EDA

PHASE 1: **FEATURE ENGINEERING**

The engagement metrics were used to create the 'category_activity_mean' feature which is what I am predicting for

cocreation_tool	project_id	user_id	content_type_name_entry	content_type_name_post	parent_content_type_name	title	categories	follow	watch
brainstorm	155	1439	0	1	project	Olli: self-driving, cognitive electric shuttle	Ground Mobility	1	0
challenge	165	75774	1	0	project	Urbanization of Mars	Mars	0	1
challenge	160	65782	1	1	project	Airbus Cargo Drone	Air Mobility	1	1

This was turned into 'total_users' a total count of users per project

This was turned into 'entry_count', a total count of entries per project

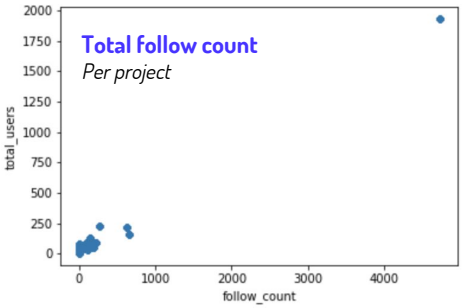
This was turned into 'post_count', a total count of posts per project

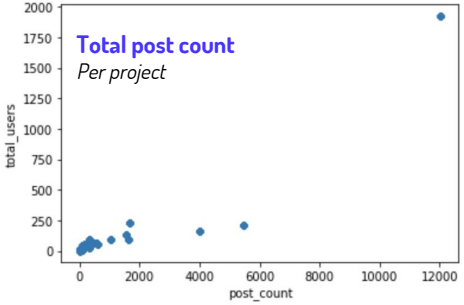
This was dummy coded and all categories were given their own column.

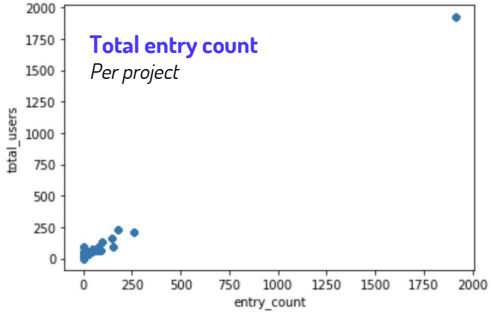
This was turned into 'follow_count', a total count of follows per project

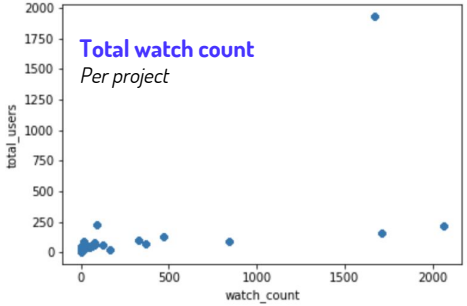
This was turned into 'watch_count', a total count of watchs per project

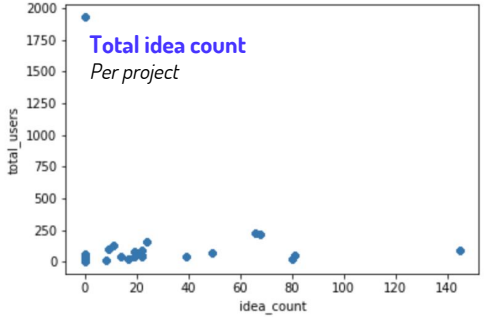
PHASE 2: **PLOTTING**

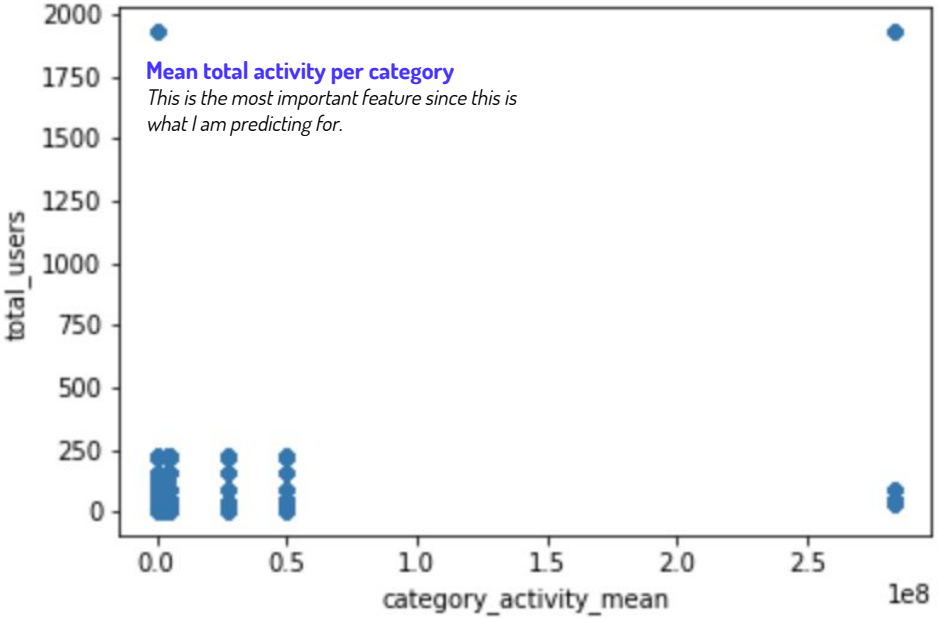




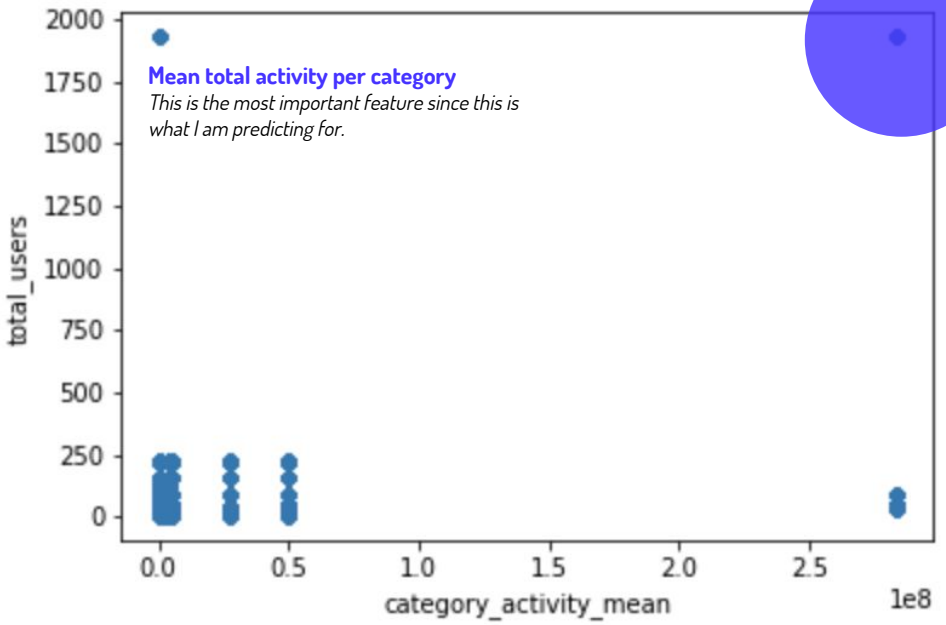








PHASE 2: **PLOTTING**



On closer inspection all of the outliers shown in the plots are from the projects in one of the categories: Mars.

Since I am trying to predict activity based on what category the project is in, I'm going to try to move forward without removing these outliers, so that I can keep that category as a part of my model.

MODELING

Modeling Round 1

BIAS VS. VARIANCE

- The training and test model MSE are not so different that there is a concern of variance.
- The difference between test and null model MSE show that bias is relatively low and not a concern.

ANALYSIS

- Other than the particularly horrible linear regression w/ cross val score, the MSEs seem to be consistent across all models with little to no improvement.
- Though I'm predicting numbers in the millions, this MSE seems overly large.

LINEAR REGRESSION W/ TRAIN TEST SPLIT	LINEAR REGRESSION W/ CROSS VAL SCORE	REGRESSION TREE	REGRESSION TREE W/ CROSS VAL SCORE	BAGGING W/ REGRESSION TREE
Training model MSE: 4370715774996228.0	**MSE:** 9.949696157450484e+32	**MSE:** 4277976634802103.0	**MSE:** 4350657404988150.5	**MSE:** 4276344882915948.5
Test model MSE: 4276341366157285.0	***Wow that is REALLY bad!***			
Null model MSE: 4633470451736635.0				

Modeling Round 2

The last round of modeling had suspiciously high MSEs, however after removing the outlier category: Mars, these MSEs are much more acceptable.

ANALYSIS

- Similar to the last round of modeling bias and variance do not seem to be a cause for concern.
- These MSE's are half of the last round of modeling and far more acceptable.

LINEAR REGRESSION W/ TRAIN TEST SPLIT

Training model MSE:
223011431199581.4

Test model MSE:
222063551042432.66

Null model MSE:
234553939504031.4

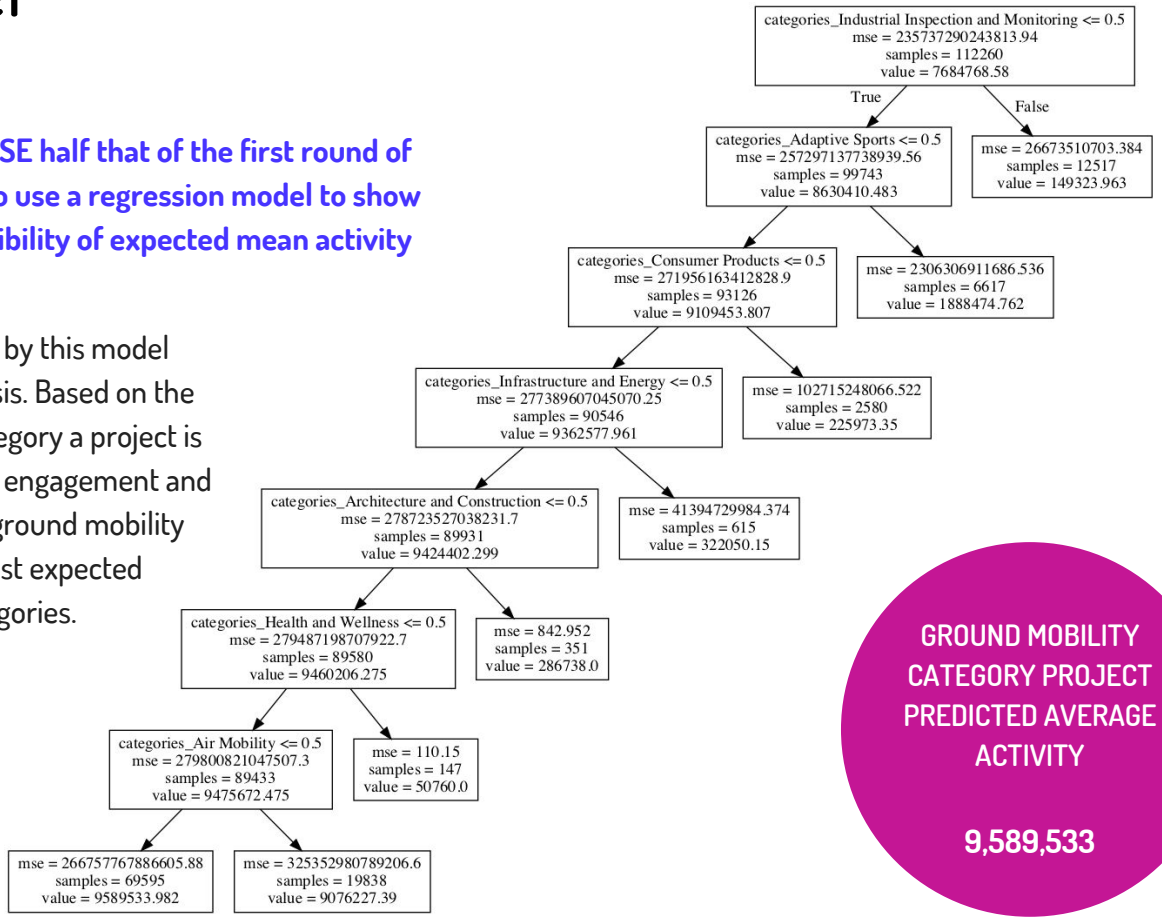
REGRESSION TREE W/ CROSS VAL SCORE

MSE:
222062762501317.06

Final Model

With the final model MSE half that of the first round of modeling, I've chosen to use a regression model to show each split for easier legibility of expected mean activity within categories.

This outcome predicted by this model proves out my hypothesis. Based on the predicted value, the category a project is in will be an indicator of engagement and the projects within the ground mobility category have the highest expected mean activity of all categories.



GROUND MOBILITY CATEGORY PROJECT PREDICTED AVERAGE ACTIVITY

9,589,533

NEXT STEPS

FINAL THOUGHTS

I am content with the MSE on the second round of modeling but the signals within the data were not as strong as I had expected. There are additional steps that can be taken to confirm and validate the conclusion.

NEXT STEPS

- Use the full data set (3+ million) rows rather than a sample
- Only use categories that have higher engagement metrics as predictors to measure how this changes the conclusion.
- Run this model on singular engagement metrics such as entries, ideas, posts and follows to get exact predicted numbers for each of these action types.

THANK YOU