Team Varun C

Final Submission by
Koushik Kulkarni, Varun, Antriksh Goel

SOLUTION APPROACH

DATA CLEANSING & PRE-PROCESSING

- Replace Missing Values
- Replace non-ASCII characters
- HTML tag extraction using Regex in Python
- Derive categorical features from text-based (non-categorical) features using libraries like NLTK, spaCy, etc
- Derive features from DATETIME type features

Insights visualization

VISUALIZATION - https://public.tableau.com/profile/varun3553#!/vizhome/PDC_Hackathon_2019/PDCHackathon2019

1. While the highest number of petition generated from the US, their success rate is quite low.
2. California is the only US State that with a high success rate along with a decent contribution of petition count.
3. Education with 31% and Tax with 33% are the categories with highest number of petitions while Education category contributes the highest number of successful petitions, which is 15.27%.
4. The petitions with score values between 2 and 3 seem to be the most successful in terms of balance in volume and success rate.
5. The probability of target being an individual and his visibility cause the success rate to go down.

Machine Learning Models

Predictive modelling

1. Extracted <mark> tags from highlight_description and highlight_ask columns
2. These columns are available in the validation dataset as well.
3. These have very linear relationship with the category
4. Trained CatBoost model to predict with 100% accuracy.

Category classification

1. Included numeric and categorical attributes
2. textual attributes not included in this version, these can be processes using text processing libraries like NLTK or Google Natural Language API or AutoML Language API to create additional attributes in next version.
3. Used CatBoost model to facilitate automatic one-hot encoding of categorical attributes
4. Class imbalance of 8
5. Final results on test data (part of train) - accuracy 90%, F1 score 88%
6. Hyperparameters could be tuned further in wider parameter space