Health Insurance Lead Prediction

Analytics Vidhya | Kaggle Competition

Miguel Angel Santana II, MBA

February 2, 2021

Introduction

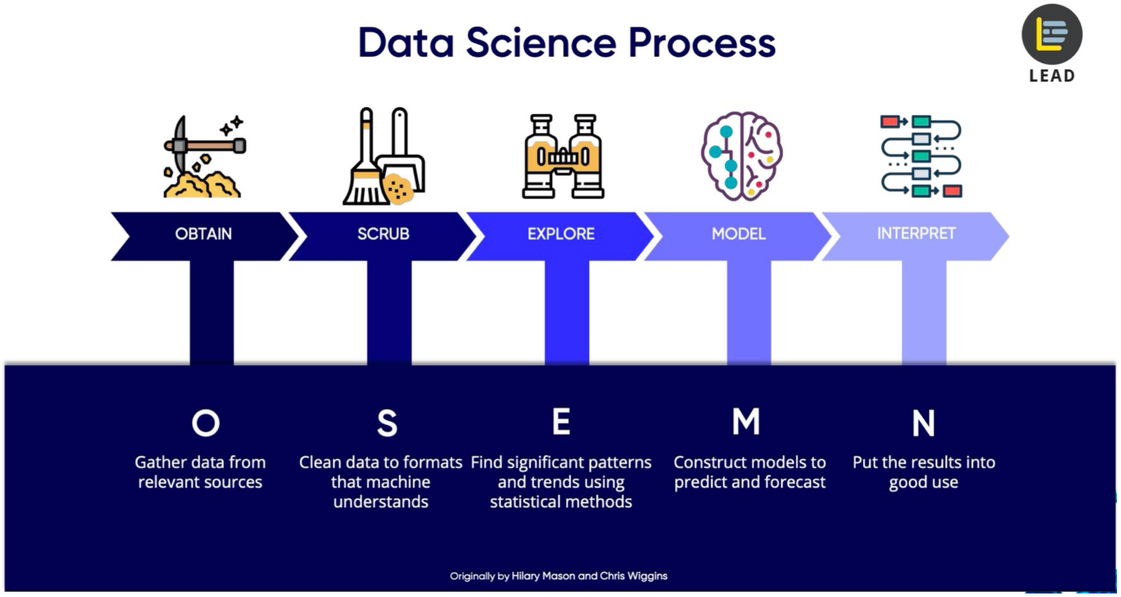
- Methodology
- Data Processing
- Model Selection
  - Validation
- Interpret Results
  - Feature Selection
- Recommendations
  - Limitations & Future Work
- Thank You

Methodology

- FinMan Company is looking to cross sell insurance products to new and existing customers.
- Insurance policies are offered to clients based on website landing and consumer choice (election to fill out forms).
- FinMan company would like to classify positive leads for outreach programs using machine learning.

Data Processing

- OSEMN Framework
- Key Decisions:
  - Filling missing values
  - Feature engineering
    - Average Age
    - Long Term Customer
    - Primary Age – Premium Factor



Data Science Process

LEAD

OBTAIN → SCRUB → EXPLORE → MODEL → INTERPRET

O Gather data from relevant sources

S Clean data to formats that machine understands

E Find significant patterns and trends using statistical methods

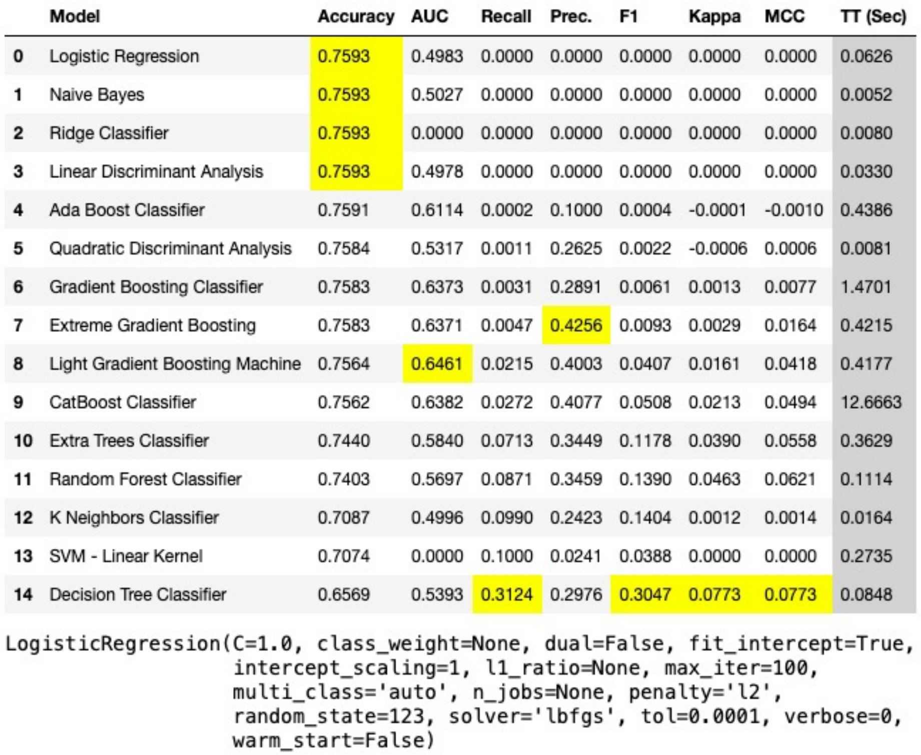
M Construct models to predict and forecast

N Put the results into good use

Originally by Hilary Mason and Chris Wiggins

Model Selection

- Preliminary models were run to narrow down our selection.
- The primary model was selected by considering AUC and overall Accuracy.
- A grid search was performed in order to improve performance.

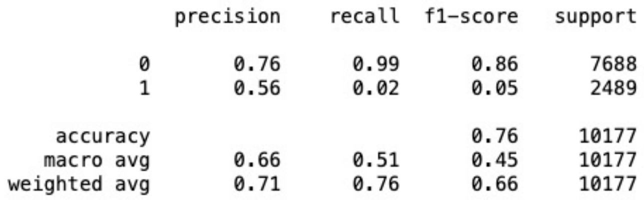


	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
0	Logistic Regression	0.7593	0.4983	0.0000	0.0000	0.0000	0.0000	0.0000	0.0626
1	Naive Bayes	0.7593	0.5027	0.0000	0.0000	0.0000	0.0000	0.0000	0.0052
2	Ridge Classifier	0.7593	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0080
3	Linear Discriminant Analysis	0.7593	0.4978	0.0000	0.0000	0.0000	0.0000	0.0000	0.0330
4	Ada Boost Classifier	0.7591	0.6114	0.0002	0.1000	0.0004	-0.0001	-0.0010	0.4386
5	Quadratic Discriminant Analysis	0.7584	0.5317	0.0011	0.2625	0.0022	-0.0006	0.0006	0.0081
6	Gradient Boosting Classifier	0.7583	0.6373	0.0031	0.2891	0.0061	0.0013	0.0077	1.4701
7	Extreme Gradient Boosting	0.7583	0.6371	0.0047	0.4256	0.0093	0.0029	0.0164	0.4215
8	Light Gradient Boosting Machine	0.7564	0.6461	0.0215	0.4003	0.0407	0.0161	0.0418	0.4177
9	CatBoost Classifier	0.7562	0.6382	0.0272	0.4077	0.0508	0.0213	0.0494	12.6663
10	Extra Trees Classifier	0.7440	0.5840	0.0713	0.3449	0.1178	0.0390	0.0558	0.3629
11	Random Forest Classifier	0.7403	0.5697	0.0871	0.3459	0.1390	0.0463	0.0621	0.1114
12	K Neighbors Classifier	0.7087	0.4996	0.0990	0.2423	0.1404	0.0012	0.0014	0.0164
13	SVM - Linear Kernel	0.7074	0.0000	0.1000	0.0241	0.0388	0.0000	0.0000	0.2735
14	Decision Tree Classifier	0.6569	0.5393	0.3124	0.2976	0.3047	0.0773	0.0773	0.0848

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=123, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```

Validation

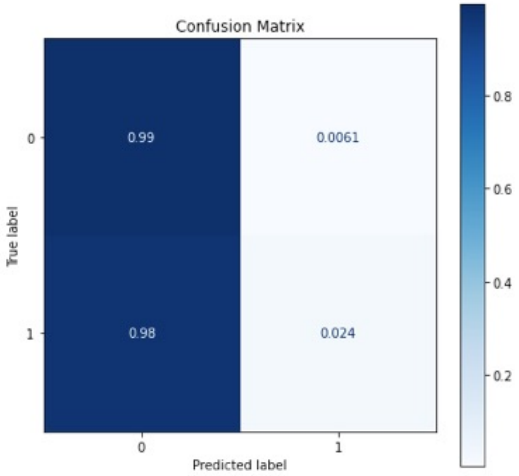
After the grid search the model scored:

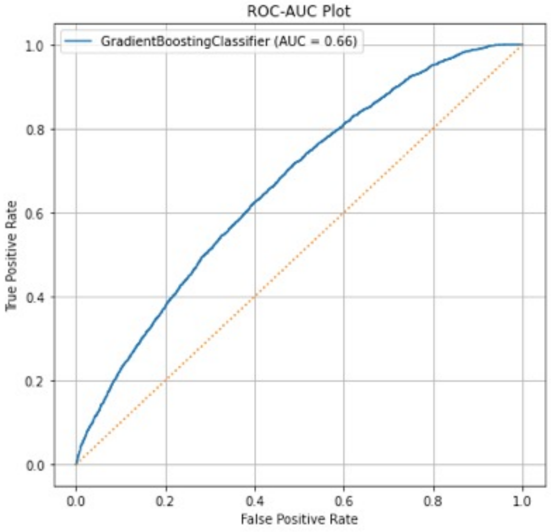


```
              precision    recall  f1-score   support

           0       0.76      0.99      0.86      7688
           1       0.56      0.02      0.05      2489

    accuracy                           0.76     10177
   macro avg       0.66      0.51      0.45     10177
weighted avg       0.71      0.76      0.66     10177
```



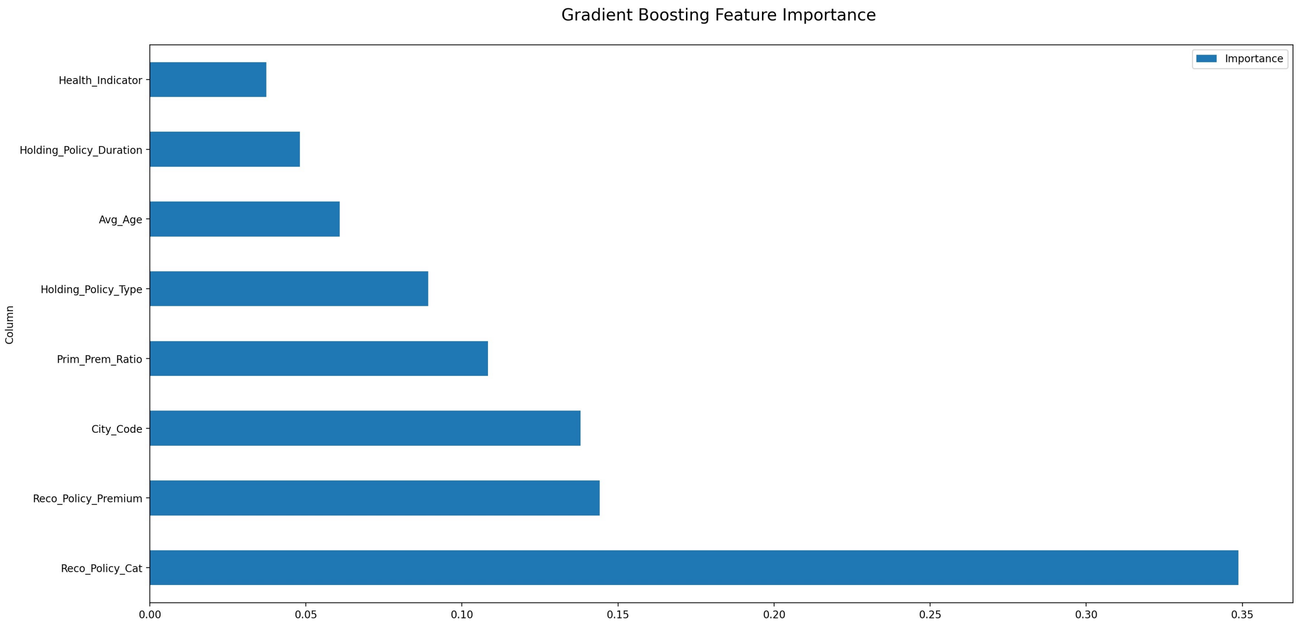


Interpret Results

Feature Selection

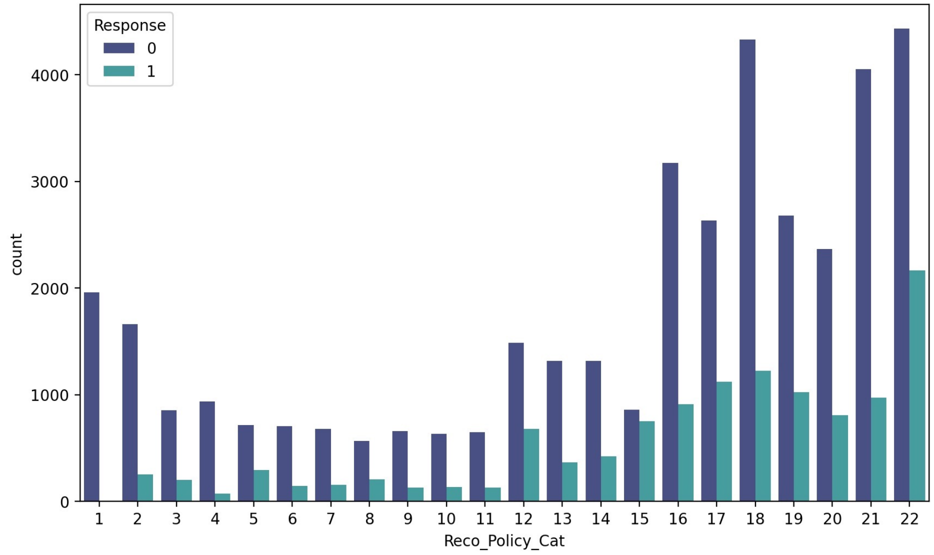
Top 3

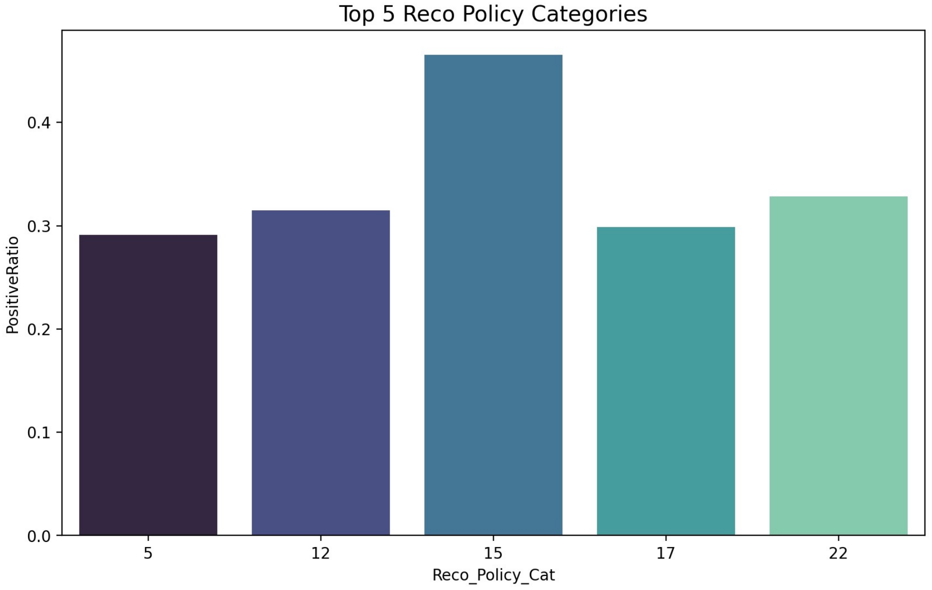
- Reco Policy Category
- Reco Policy Premium
- City Code



Feature Selection cont.

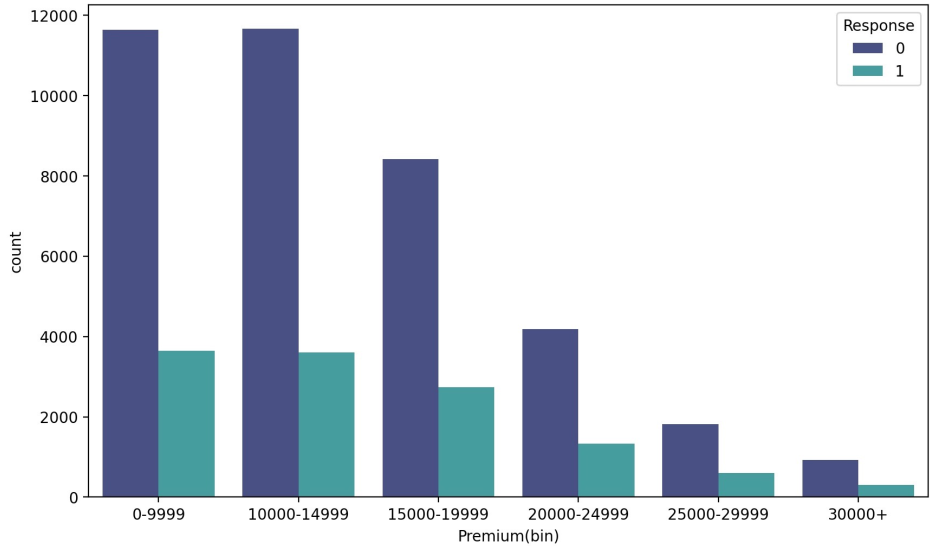
- Reco Policy Category
  - Positive to total response ratio
  - Target Policy Categories
    - 15
    - 22

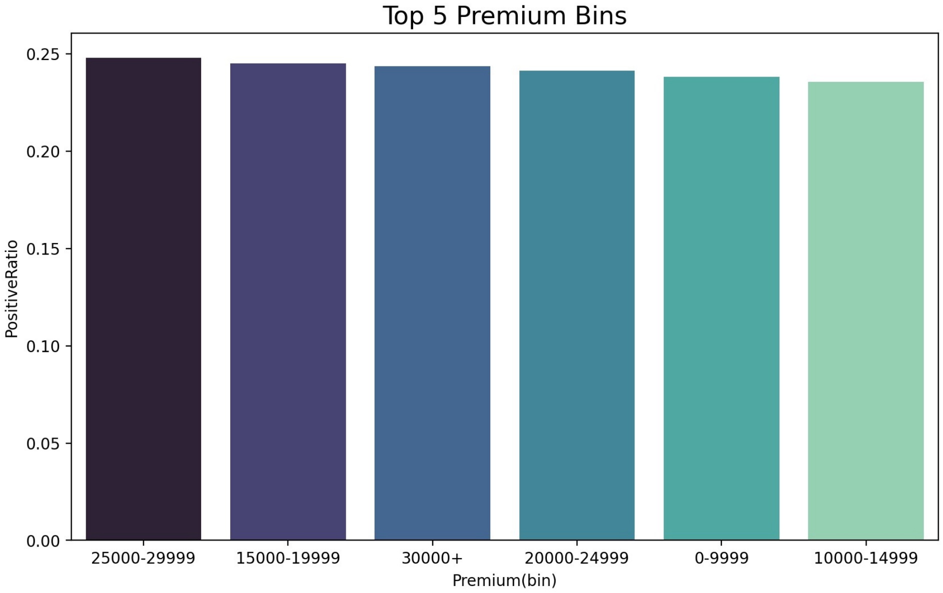




Feature Selection cont.

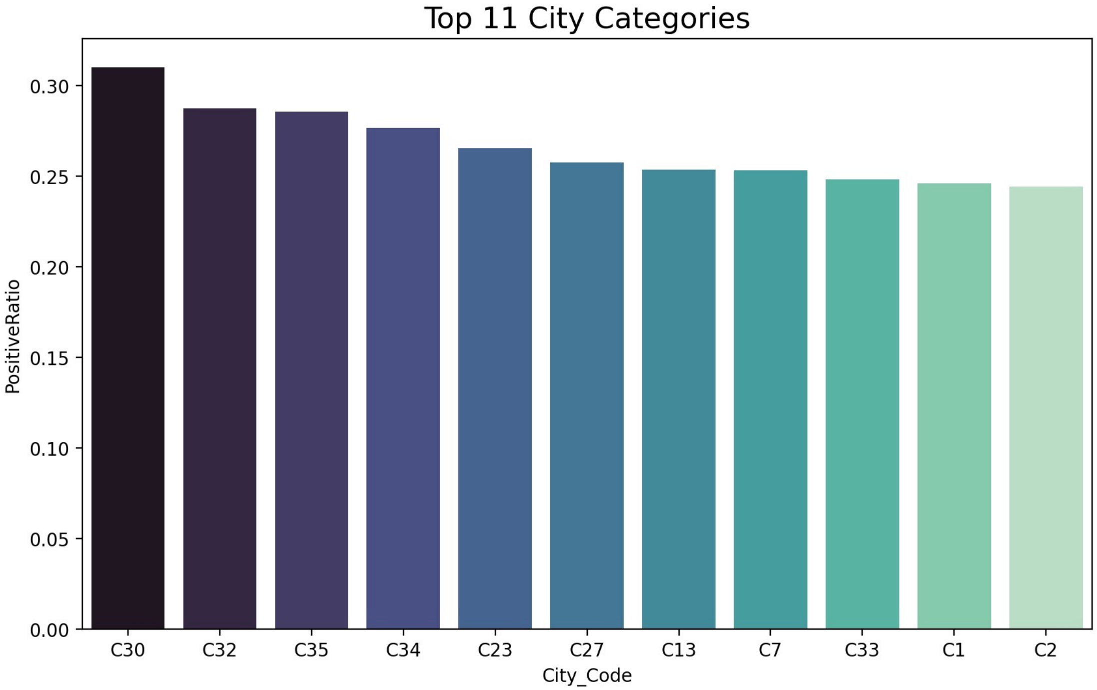
- Reco Policy Premium
  - Positive to total response ratio
  - Target Clients with Premiums Between:
    - $15,000 – $19,999





Feature Selection cont.

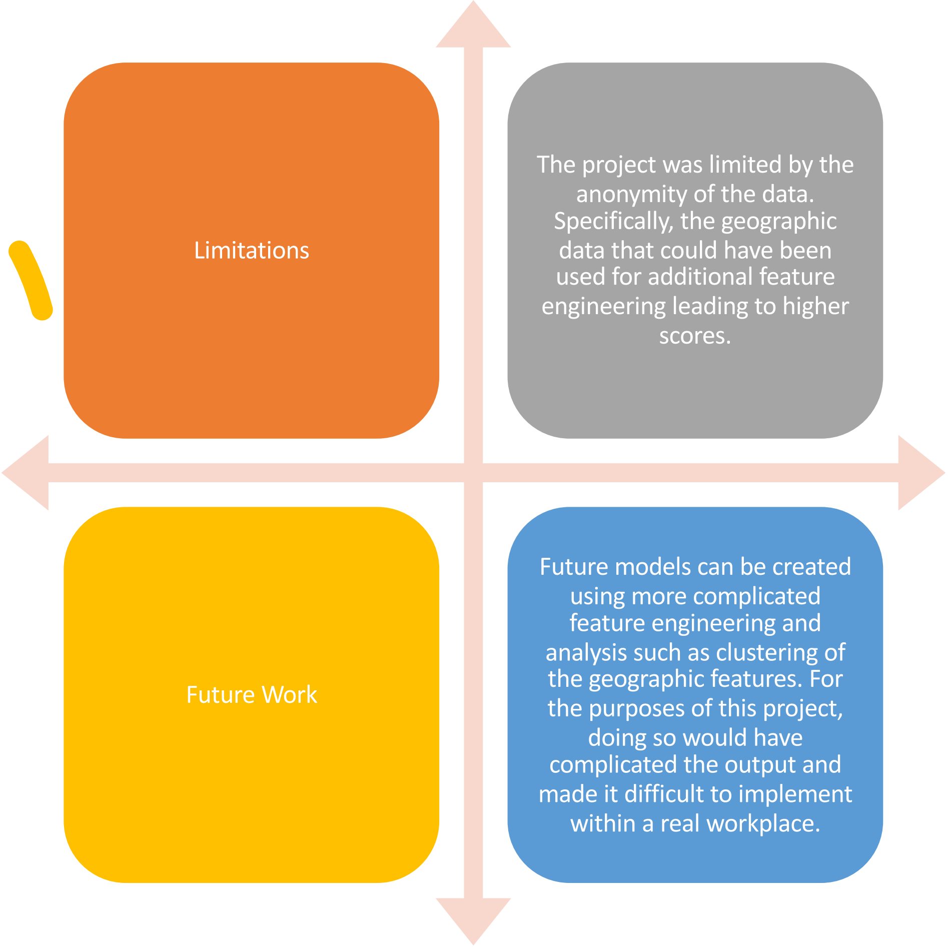
- City Code
  - Positive to total response ratio
  - Target Clients Living in:
    - C1, C2, C13, C23
    - C1 & C2 have a similar response ratio but a much larger client volume.



Recommendations

The model's top 3 features were: Reco Policy Category, Reco Policy Premium and City Code.

- It is recommended to focus on clients in/with:
  - Reco Policy Categories: 15, 22
  - Reco Policy Premiums: Between $15,000 - $19,999.
  - City Codes: C1, C2, C13, C23



Limitations

The project was limited by the anonymity of the data. Specifically, the geographic data that could have been used for additional feature engineering leading to higher scores.

Future Work

Future models can be created using more complicated feature engineering and analysis such as clustering of the geographic features. For the purposes of this project, doing so would have complicated the output and made it difficult to implement within a real workplace.

THANK YOU!

Questions?

Miguel Santana

Contact: santana2.miguel@gmail.com

Additional projects can be found on Github.

Username: miguelangelsantana