Presence of Opioids in Wastewater

Drew Gobbi, Olivia Wang, Maddie Warndorf

PU1

Tempe, AZ - Classification and Feature Importance Experimentation

- Attempted classification approach for tracts with a greater # of EMS calls than dataset median
- Using Logistic Regression, find that model performs poorly and determines that essentially meaningless (such as the tract ID, and the land area, shape area) are important

```
In [ ]: 1 print(classification_report(y_test, pred))
```

	precision	recall	f1-score	support
0	0.68	0.80	0.73	55
1	0.52	0.36	0.43	33
micro avg	0.64	0.64	0.64	88
macro avg	0.60	0.58	0.58	88
weighted avg	0.62	0.64	0.62	88

```
In [ ]: 1 confusion_matrix(y_test, pred)

Out[ ]: array([[44, 11],
               [21, 12]])
```

Confusion matrix

Tempe, AZ - Classification and Feature Importance Experimentation

- Using Random Forest, find similarly poor performance (top 5 features below)
- Issues may be model specific, but will attempt to standardize features and experiment further

	precision	recall	f1-score	support
0	0.71	0.87	0.78	55
1	0.65	0.39	0.49	33
micro avg	0.69	0.69	0.69	88
macro avg	0.68	0.63	0.64	88
weighted avg	0.68	0.69	0.67	88

```
[[48  7]
 [20 13]]
```

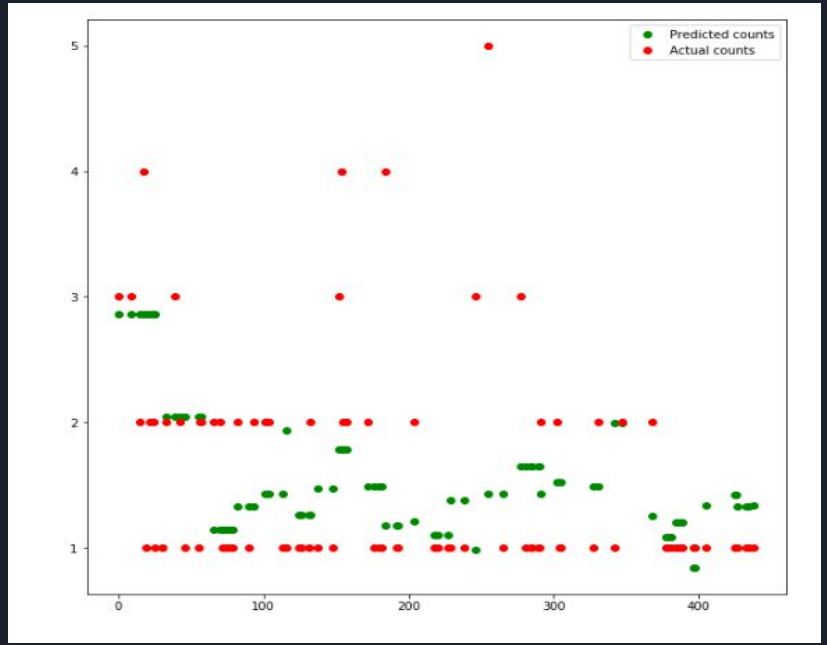
Confusion matrix

	feature	importance
3	Incident_Date	0.116578
7	Longitude_Random	0.116201
2	OBJECTID_x	0.114190
4	Month	0.109508
6	Latitude_Random	0.109463

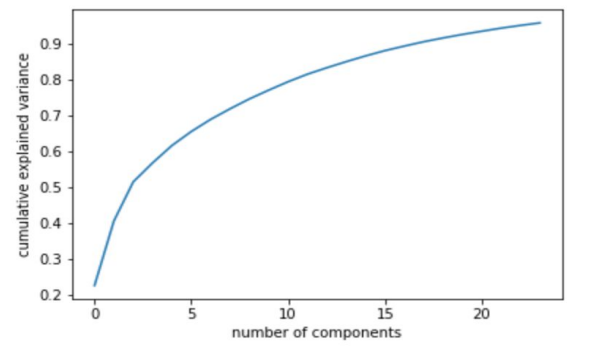
Recreated the Poisson Model for Tempe, AZ

- Used targeted features based on literature and hypotheses (demographics and facilities)
- Found that percent of the college population and the concentrations of naloxone and drug drop off were significant
- Model deviance still suggests poor goodness of fit
- Appear to be scale and colinearity issues impacting performance

Predicted vs Actual EMS Call Counts/tract/month



Analyzing Tempe EMS Calls with PCA



- Standardized the dataset to (0,1) range and ran a PCA → found n=24 components
- Three components were found to be statistically significant
- Need help with executing next steps:
  - Completing loading analysis to determine key factors
  - Inverse transform of data to make predictions and interpret coefficients correctly

Generalized Linear Model Regression Results

Dep. Variable:	ems_calls	No. Observations:	352
Model:	GLM	Df Residuals:	328
Model Family:	Poisson	Df Model:	23
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-482.56
Date:	Tue, 04 Feb 2020	Deviance:	165.34
Time:	20:29:03	Pearson chi2:	226.
No. Iterations:	4	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
x1	0.0007	0.005	0.130	0.897	-0.009	0.011
x2	-0.0155	0.006	-2.461	0.014	-0.028	-0.003
x3	-0.0005	0.007	-0.073	0.942	-0.015	0.014
x4	0.0553	0.009	5.885	0.000	0.037	0.074
x5	-0.0254	0.011	-2.268	0.023	-0.047	-0.003
x6	-0.0042	0.012	-0.340	0.734	-0.028	0.020
x7	0.0056	0.014	0.392	0.695	-0.023	0.034
x8	-0.0048	0.016	-0.304	0.761	-0.036	0.026
x9	-0.0036	0.015	-0.237	0.813	-0.033	0.026
x10	0.0195	0.015	1.283	0.199	-0.010	0.049
x11	0.0068	0.017	0.407	0.684	-0.026	0.039
x12	0.0121	0.019	0.646	0.518	-0.025	0.049
x13	0.0215	0.022	0.997	0.319	-0.021	0.064
x14	-0.0273	0.020	-1.343	0.179	-0.067	0.013
x15	-0.0275	0.022	-1.268	0.205	-0.070	0.015
x16	-0.0095	0.022	-0.440	0.660	-0.052	0.033
x17	-0.0077	0.023	-0.336	0.737	-0.053	0.037
x18	0.0094	0.023	0.407	0.684	-0.036	0.055
x19	0.0131	0.026	0.495	0.621	-0.039	0.065
x20	0.0233	0.027	0.855	0.393	-0.030	0.077
x21	0.0313	0.026	1.194	0.233	-0.020	0.083
x22	-0.0191	0.027	-0.704	0.481	-0.072	0.034
x23	0.0071	0.031	0.225	0.822	-0.054	0.068
x24	0.0515	0.030	1.725	0.084	-0.007	0.110

Adjusting Cincinnati, OH EMS Data vs. Census Data

Correlation Matrix (shown on the right)

- Some of the medical facility attributes are highly correlated so we got rid of some of them and reran poisson model and logit model on the rest of the attributes and EMS calls
- Because HospitalGeneralCount2mi and SubstanceAbuse2mil are highly correlated we tried getting rid of either of them and the significant attributes came back differently

	CensusTract	HospitalsGeneralCount2mi	HospitalsPsychCount2mi	SurgicalCentersCount2mi	UrgentCareCount2mi	WomensClinic
CensusTract	1.000000	-0.467770	0.113171	-0.338477	-0.231234	
HospitalsGeneralCount2mi	-0.467770	1.000000	-0.168916	0.599794	0.318476	
HospitalsPsychCount2mi	0.113171	-0.168916	1.000000	0.088782	0.218105	
SurgicalCentersCount2mi	-0.338477	0.599794	0.088782	1.000000	0.613852	
UrgentCareCount2mi	-0.231234	0.318476	0.218105	0.613852	1.000000	
WomensClinicsCount2mi	-0.465796	0.834486	-0.342838	0.303941	0.112762	
PainManagementCount2mi	-0.401127	0.582379	-0.175815	0.245675	0.393595	
PhysicalTherapyCount2mi	-0.424851	0.806142	-0.084581	0.670437	0.475266	
MentalFacilitiesCount2mi	-0.485608	0.897106	-0.235316	0.603392	0.255753	
SubstanceAbuseCount2mi	-0.481727	0.936103	-0.325055	0.540802	0.251382	
OtherMedicalFacilitesCount2mi	-0.549316	0.887057	-0.276280	0.466488	0.285536	
NursingAssistedLiveCount2mi	0.057260	0.231820	0.069165	0.443496	0.285741	
HospiceCount2mi	-0.227312	0.328493	0.376019	0.299581	0.303395	
HospitalsChildCount2mi	-0.413186	0.909247	-0.277820	0.587578	0.175458	
ChildFacilitiesCount2mi	-0.267150	0.051810	-0.143782	-0.117629	0.079678	
PharmCount2mi	-0.457994	0.857705	-0.252529	0.433753	0.433896	
DrugDropCount2mi	-0.331989	0.580334	-0.351102	0.337106	0.142843	
NaloxoneDistribCount2mi	-0.350512	0.618917	-0.209786	0.259367	0.373292	

Poisson Model with GeneralHospital2mil Left Out

Generalized Linear Model Regression Results

Dep. Variable:	ems_call	**No. Observations:**	3245
Model:	GLM	**Df Residuals:**	3106
Model Family:	Poisson	**Df Model:**	138
Link Function:	log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-21510.
Date:	Wed, 05 Feb 2020	**Deviance:**	26593.
Time:	19:29:01	**Pearson chi2:**	2.69e+04
No. Iterations:	100	**Covariance Type:**	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
HospitalsPsychCount2mi	-0.0757	0.036	-2.109	0.035	-0.146	-0.005
SurgicalCentersCount2mi	0.1252	0.023	5.386	0.000	0.080	0.171
UrgentCareCount2mi	-0.4711	0.019	-24.427	0.000	-0.509	-0.433
PainManagementCount2mi	0.5368	0.046	11.787	0.000	0.448	0.626
MentalFacilitiesCount2mi	0.0651	0.016	3.961	0.000	0.033	0.097
SubstanceAbuseCount2mi	0.0396	0.019	2.042	0.041	0.002	0.078
NursingAssistedLiveCount2mi	0.0561	0.004	12.714	0.000	0.047	0.065
HospiceCount2mi	-0.0630	0.040	-1.582	0.114	-0.141	0.015
ChildFacilitiesCount2mi	0.4102	0.022	18.305	0.000	0.366	0.454
DrugDropCount2mi	-0.1991	0.018	-11.168	0.000	-0.234	-0.164
NaloxoneDistribCount2mi	0.0441	0.004	12.343	0.000	0.037	0.051
Median_Age_ACS_13_17	-0.0009	0.001	-1.426	0.154	-0.002	0.000
pct_College_ACS_13_17	-6.334e-05	0.000	-0.276	0.783	-0.001	0.000
pct_Not_HS_Grad_ACS_13_17	2.876e-05	0.001	0.052	0.958	-0.001	0.001
pct_NoHealthIns_65P_ACS_13_17	-0.0007	0.001	-0.693	0.489	-0.003	0.001
pct_NH_White_alone_ACS_13_17	-3.005e-05	0.000	-0.213	0.831	-0.000	0.000
pct_NH_AIAN_alone_ACS_13_17	0.0202	0.005	4.439	0.000	0.011	0.029
pct_Males_ACS_13_17	0.0012	0.001	2.157	0.031	0.000	0.002
pct_Pop_18_24_ACS_13_17	-0.0006	0.000	-1.815	0.069	-0.001	4.46e-05
pct_Prs_Blw_Pov_Lev_ACS_13_17	0.0004	0.000	1.027	0.304	-0.000	0.001
pct_One_Health_Ins_ACS_13_17	0.0005	0.000	1.036	0.300	-0.000	0.001
pct_Civ_unemp_16_24_ACS_13_17	-0.0005	0.000	-2.325	0.020	-0.001	-8.34e-05
Civ_labor_25_44_ACS_13_17	-6.294e-06	8.9e-06	-0.707	0.479	-2.37e-05	1.11e-05
pct_Civ_unemp_45_64_ACS_13_17	0.0002	0.000	0.410	0.682	-0.001	0.001
pct_PUB_ASST_INC_ACS_13_17	-0.0018	0.001	-2.488	0.013	-0.003	-0.000

Poisson Model with SubstanceAbuse2mi Left Out

Generalized Linear Model Regression Results

Dep. Variable:	ems_call	**No. Observations:**	3248
Model:	GLM	**Df Residuals:**	3109
Model Family:	Poisson	**Df Model:**	138
Link Function:	log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-21745.
Date:	Wed, 05 Feb 2020	**Deviance:**	27016.
Time:	19:33:52	**Pearson chi2:**	2.74e+04
No. Iterations:	100	**Covariance Type:**	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
HospitalsGeneralCount2mi	0.2231	0.015	15.084	0.000	0.194	0.252
HospitalsPsychCount2mi	-0.1587	0.038	-4.217	0.000	-0.232	-0.085
SurgicalCentersCount2mi	0.1002	0.021	4.710	0.000	0.059	0.142
UrgentCareCount2mi	-0.4025	0.018	-22.496	0.000	-0.438	-0.367
PainManagementCount2mi	0.4359	0.035	12.523	0.000	0.368	0.504
MentalFacilitiesCount2mi	-0.0817	0.023	-3.606	0.000	-0.126	-0.037
NursingAssistedLiveCount2mi	0.0330	0.004	7.605	0.000	0.024	0.041
HospiceCount2mi	-0.0830	0.037	-2.219	0.026	-0.156	-0.010
ChildFacilitiesCount2mi	0.4509	0.027	16.674	0.000	0.398	0.504
DrugDropCount2mi	-0.1273	0.017	-7.681	0.000	-0.160	-0.095
NaloxoneDistribCount2mi	0.0157	0.004	4.269	0.000	0.009	0.023
Median_Age_ACS_13_17	-0.0002	0.001	-0.364	0.716	-0.001	0.001
pct_College_ACS_13_17	-0.0003	0.000	-1.158	0.247	-0.001	0.000
pct_Not_HS_Grad_ACS_13_17	0.0007	0.001	1.324	0.185	-0.000	0.002
pct_NoHealthIns_65P_ACS_13_17	0.0015	0.001	1.441	0.150	-0.001	0.004
pct_NH_White_alone_ACS_13_17	-0.0002	0.000	-1.449	0.147	-0.000	7.27e-05
pct_NH_AIAN_alone_ACS_13_17	-0.0207	0.005	-4.157	0.000	-0.030	-0.011
pct_Males_ACS_13_17	0.0012	0.001	2.156	0.031	0.000	0.002
pct_Pop_18_24_ACS_13_17	-0.0003	0.000	-0.788	0.431	-0.001	0.000
pct_Prs_Blw_Pov_Lev_ACS_13_17	-0.0014	0.000	-4.023	0.000	-0.002	-0.001
pct_One_Health_Ins_ACS_13_17	0.0014	0.000	3.077	0.002	0.001	0.002
pct_Civ_unemp_16_24_ACS_13_17	9.701e-05	0.000	0.428	0.668	-0.000	0.001
Civ_labor_25_44_ACS_13_17	8.583e-07	8.92e-06	0.096	0.923	-1.66e-05	1.83e-05
pct_Civ_unemp_45_64_ACS_13_17	-0.0004	0.000	-0.912	0.362	-0.001	0.000
pct_PUB_ASST_INC_ACS_13_17	0.0010	0.001	1.475	0.140	-0.000	0.002

Additional Tests on Cincinnati

Logit Model Results are still not very good

Logit Regression Results

Dep. Variable:	ems_flag	**No. Observations:**	4058
Model:	Logit	**Df Residuals:**	4034
Method:	MLE	**Df Model:**	23
Date:	Tue, 04 Feb 2020	**Pseudo R-squ.:**	0.1946
Time:	20:17:25	**Log-Likelihood:**	-2259.5
converged:	True	**LL-Null:**	-2805.3
		LLR p-value:	4.053e-216

ARCOS Dataset

A new dataset we found is the Automated Reports and Consolidated Ordering System (ARCOS) from the DEA that collects controlled substances transaction from manufacturers and distributors in the United States.

The reporting periods that we will pull from this data set are from 1/01/2018 - 6/30/19. There are 6 reports in each reporting period. We will use Reports 1-5 to additional context to test the identified factors in our hypothesis as being correlated with a measure of opioid use (total # grams sold).

https://www.deadiversion.usdoj.gov/arcos/retail_drug_summary/

We are almost complete with scraping the data from the PDFs.

Next Steps:

- For Tempe, AZ EMS Calls (Drew to meet with Dr. B)
  - Factor Analysis/Loadings for Tempe, AZ EMS Calls
  - Get Predictions (how to transform y var)
- Begin analysis work on PNML data
- Begin analysis on ARCOS dataset