Data & Visual Analytics Project Progress Report

Nopioid: Targeting High-Residual Opioid Prescription Medical Providers

Matt Liedtke, Prasanta Lenka, Supriya Belaguru Suresha, Pallavi Joshi, Zach Olivier

Introduction-Motivation

The Center for Disease Control and Prevention (CDC) estimates that around 68% of the more than 70,200 drug overdose deaths in 2017 involved an opioid. According to their research, the number of overdose deaths involving opioids was 6 times higher in 2017 than in 1999[20]. Studies estimate that opioid related harm costs the US about $78 billion annually[24]. Therefore, it is important to minimize the risks involved in opioid related medication.

Directly targeting over-scripting practices will positively affect one of the largest factors in the ignition of the opioid crisis – opioid prescriptions.

Problem Statement

Determine which US medical providers are writing above expected levels of opioid prescriptions. Provide insights into what factors are driving provider prescription rates.

The National Institute of Drug Abuse estimates roughly 21%-29% of patients prescribed opioids for chronic pain misuse the opioids, and around 80% of people who use heroin first misused prescription opioids. Identifying providers who write higher than normal levels of opioid prescriptions relative to their attributes will help efforts in education, awareness, and audits to help solve the opioid crisis at its core[1].

Survey:

Studies estimate that 115 people die in the US every day due to opioid overdose[25] which is closely associated with the Potentially Inappropriate Prescribing (PIP) practices of opioid[23]. Therefore, the concerned authorities need to identify where resources should be targeted to reduce opioid overdose related risks. The Department of Health and Human Services conducts audits on medical practices to analyze opioid prescribing and dispensing practices. Audits are typically very costly and inefficient. Existing visualizations facilitate in viewing trends but fail to identify over prescribing providers. Currently, most analytics and modeling efforts have focused on predicting patient level risk to opioid addiction[2], ignoring the role of providers in the crisis.

Our project aims to identify providers whose rate of opioid prescription is above the expected levels based on their attributes. Our dataset to be analyzed has many issues including missing data values and high cardinality categorical features. To deal with these issues, we will be using vTreat, a data processor for predictive modeling[6]. In this framework, the missing values are replaced with reasonable values and further supported by dummy variable columns. Also, high cardinal features are encoded in a statistically sound manner which, however, are not substitutes for domain knowledge. Features such as age, gender, and prescription history are additional domain-focused variables to consider including in our analysis[7].

To understand the relationships between provider attributes and prescription rates, exploratory data analysis needs to be performed. This includes expanding on the top 10% of opioid prescribers[8] with a more detailed Hierarchical Latent Class Clustering Analysis[9] which provides meta-cluster groups based on categorical variables. Many options are available to solve categorical variable clustering including variants of K-Means, K-Modes[10] and K-Prototypes[11]. Further, if our dataset contains both categorical and numerical variables, clustering becomes more difficult to extract insights. Probabilistic based clustering algorithms like t-SNE[12] and DBSCAN[13] are options to counteract traditional clustering issues with our dataset.

To correctly interpret our prediction model's output, model interpretation methods like SHAP[15] and LIME[16] will be used. These will allow us to determine the attributes contributing to our model's prediction. Both SHAP and LIME values aim to quantify the non-linear relationship between variables in a fitted machine learning model on the *observation level*. This allows us to show the features (by provider) which influenced our model's prediction.

To visualize our model's output and provide insights into the "where" and "why" of our results in an interactive interface, our team plans to make use of the visualization toolkits such as Tableau and ECharts. ECharts framework allows for multi-thread rendering of our results which provides efficient visualization of complex visuals sourced from large datasets. Further, the upfront coding of these visualizations is handled with only a few lines of code, a huge advantage over D3.js and other web-based visualization frameworks[17].

Proposal methods

***Intuition**:* Existing public, government visualizations show *opioid beneficiary* trends at the county and state level. It can be seen that opioid prescribing rates are broadly decreasing across the US. However, these visualizations fail to offer insight at the *provider level*. Our team will use publicly available data to identify over-prescribing providers. We will do so with the following innovations:

1) Advanced modeling, analysis and visualization of prescription rates, prescription length at the provider level
2) Utilization of SHAP values to offer on-demand insight on what features are particularly impactful in predicting a particular provider's opioid prescribing - essentially opening the complex "black box" model for clear insights
3) Clustering, analysis, and visualization of the high residual providers by key variables, allowing end users to interactively explore our results

Description:

The primary dataset is called "2017 Provider Summary Data File" from the Center for Medicare and Medicare Services. It is 1.1 million rows, by unique provider, that contains opioid prescription count and rate, as well as features that describe the providers' Medicare beneficiaries. We further enhanced this with provider attribute data (graduation year, number of individuals in practice), as well as zip-code data (population density, rural/urban classification).

It is notable that our data only captures a given provider's Medicare patient population. The majority of the Medicare population qualifies when they reach an age of 65 years, although some patients qualify for Medicare before 65 if they have a disability.

Two targets were identified to model:

Model Target	**Definition**	**Intuition**
Opioid prescription rate	total opioid prescriptions divided by total prescriptions	Captures providers who write opioids less often
Opioid days supply	The aggregate number of day's supply for opioid drugs	Captures providers who write excessively long orders or at high frequency

Once we have an accurate model for each target, we can then append predictions to all 1.1 million providers in our dataset, and compare our predictions to the actual values. Providers who have high positive residuals (actuals - predictions) will be our focus - these are the providers who are over-scripting.

Visualization and analysis on these outliers will be our main product deliverable, showing insight into who, where, and why these providers are selected as outliers.

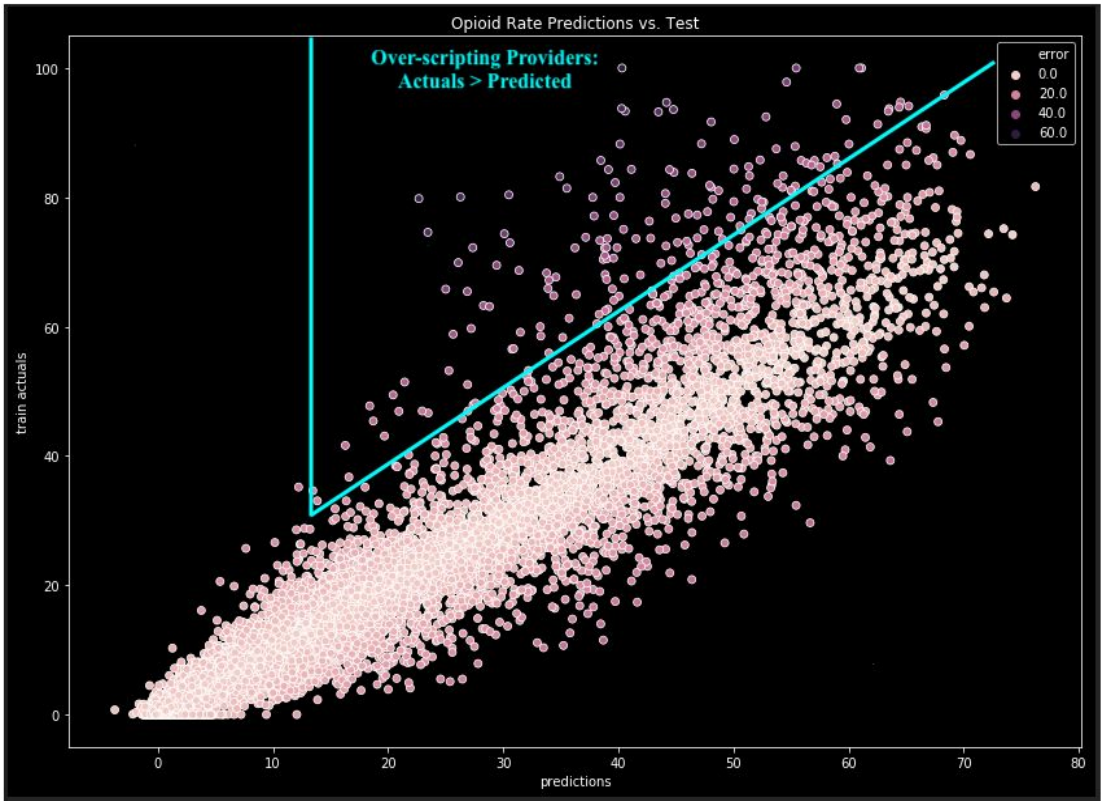
Experiments/Evaluation:

Model Development / Results:

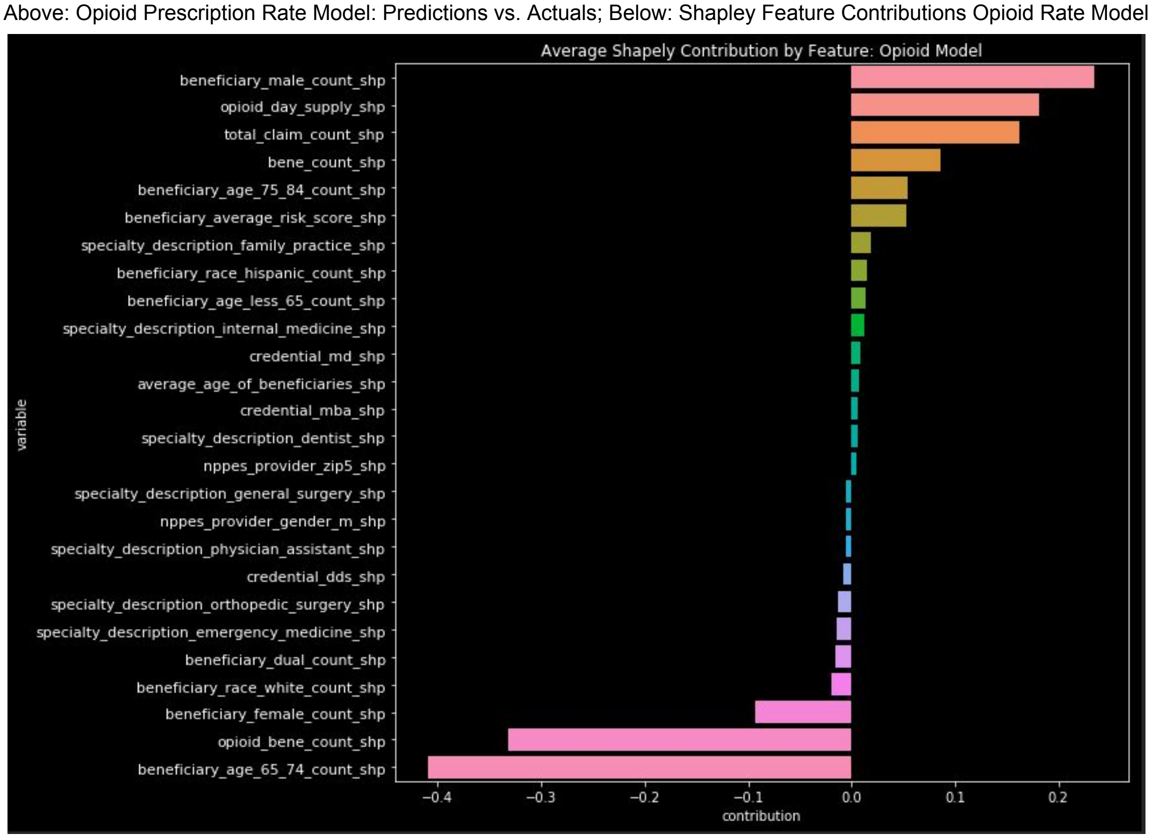
We modeled both targets using Gradient Boosting Regression with Bayesian hyper-parameter tuning. Results are promising as both models had a low mean absolute error at around 1.5%. Having a strong model allows us to hone in on providers who are truly outliers, rather than residuals due to model skill.

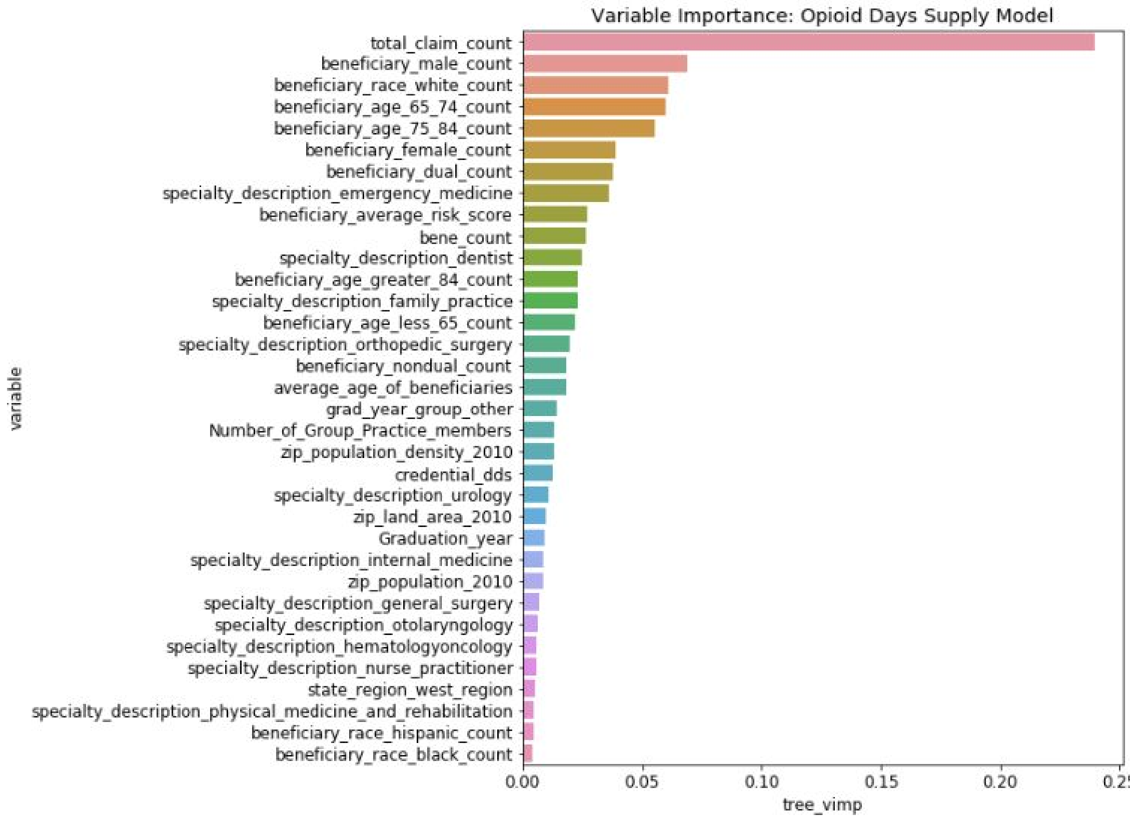
With this model we can extract global feature importance - which shows and overall view into why our model's predictions. Output of these global features will drive our follow up analysis on the high-residual providers including clustering on the key features identified by both models.

Utilizing SHAPLEY values[15], we are also able to show provider level model explanation of the predictions. This will allow model inference and feature importance at the provider level, not just at the entire training sample level. SHAPLEY contributions will be an integral part of our final visualization output and will help end users understand the "why" behind our model's predictions for individual providers.



Above: Opioid Prescription Rate Model: Predictions vs. Actuals; Below: Shapley Feature Contributions Opioid Rate Model

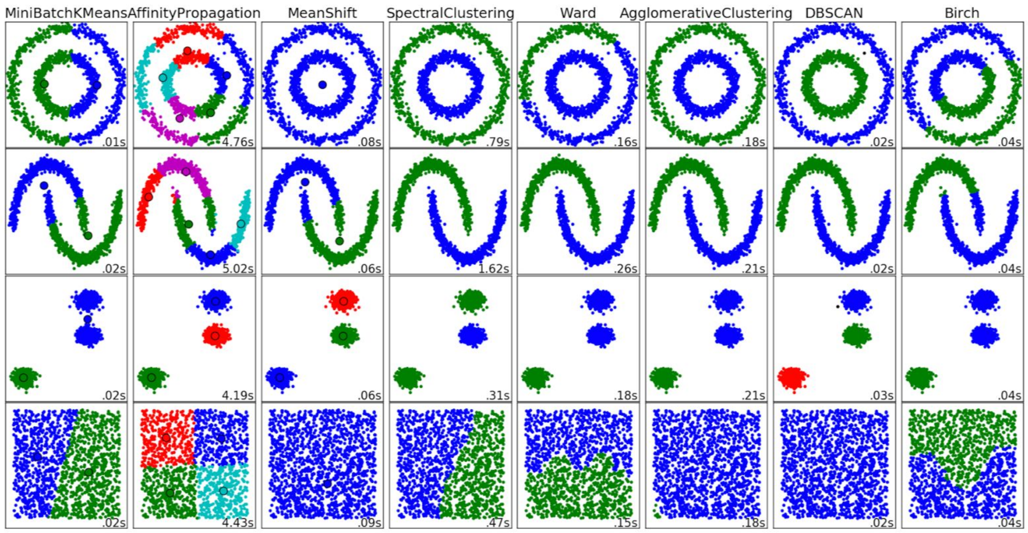




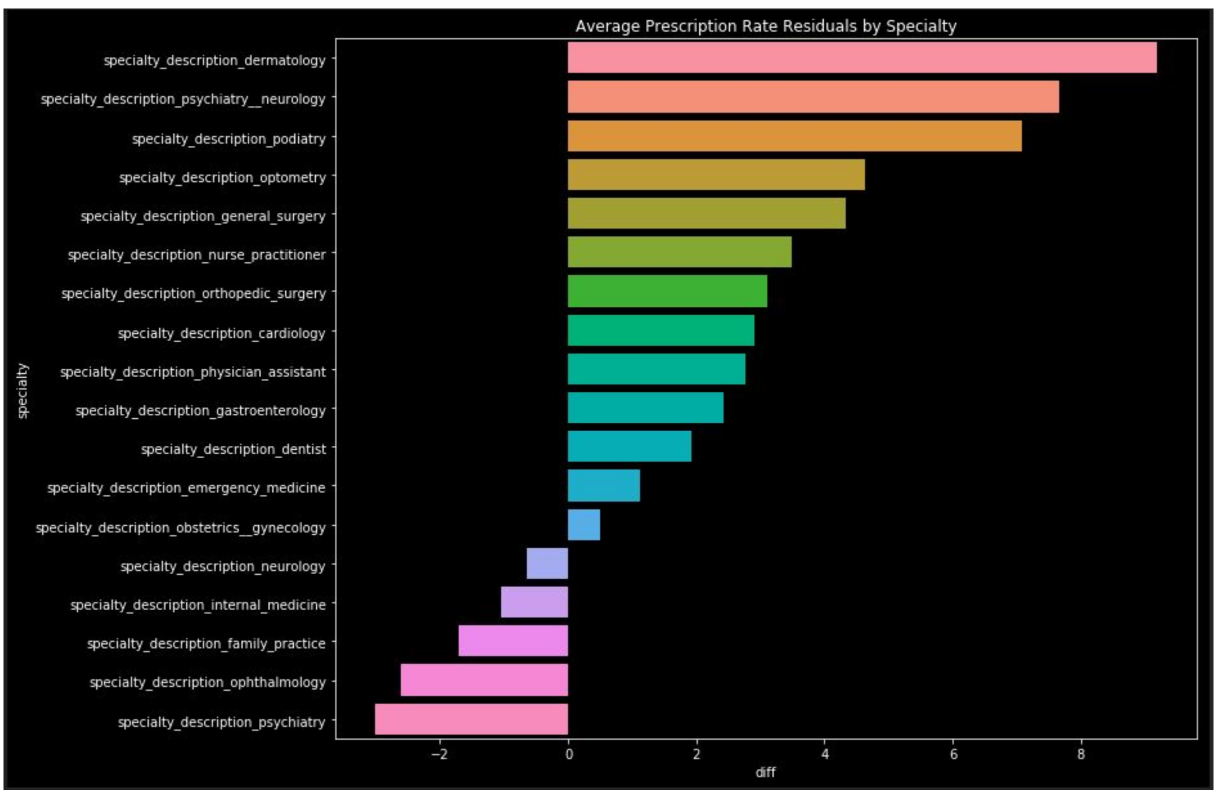
<u>Outlier Clustering</u>

After identifying our high outliers providers, we plan to implement cutting edge clustering algorithms such as DBSCAN to organize our clusters across a select key variables. These key variables include - provider credentials, geography, provider specialty and more. Clustering will allow us to segment our model selected outliers into more interpretable groups and will aid in end user interpretation.

Initial results show there is variability among these features throughout the dataset, but further insights will come from clustering the outlier providers. DBSCAN will allow us to handle the high-cardinality of our most interesting features, as well as perform outlier detection (outliers of the high residual providers)



DBSCAN in Action: Finding Clusters in high dimensional, non-parametric feature spaces models identify

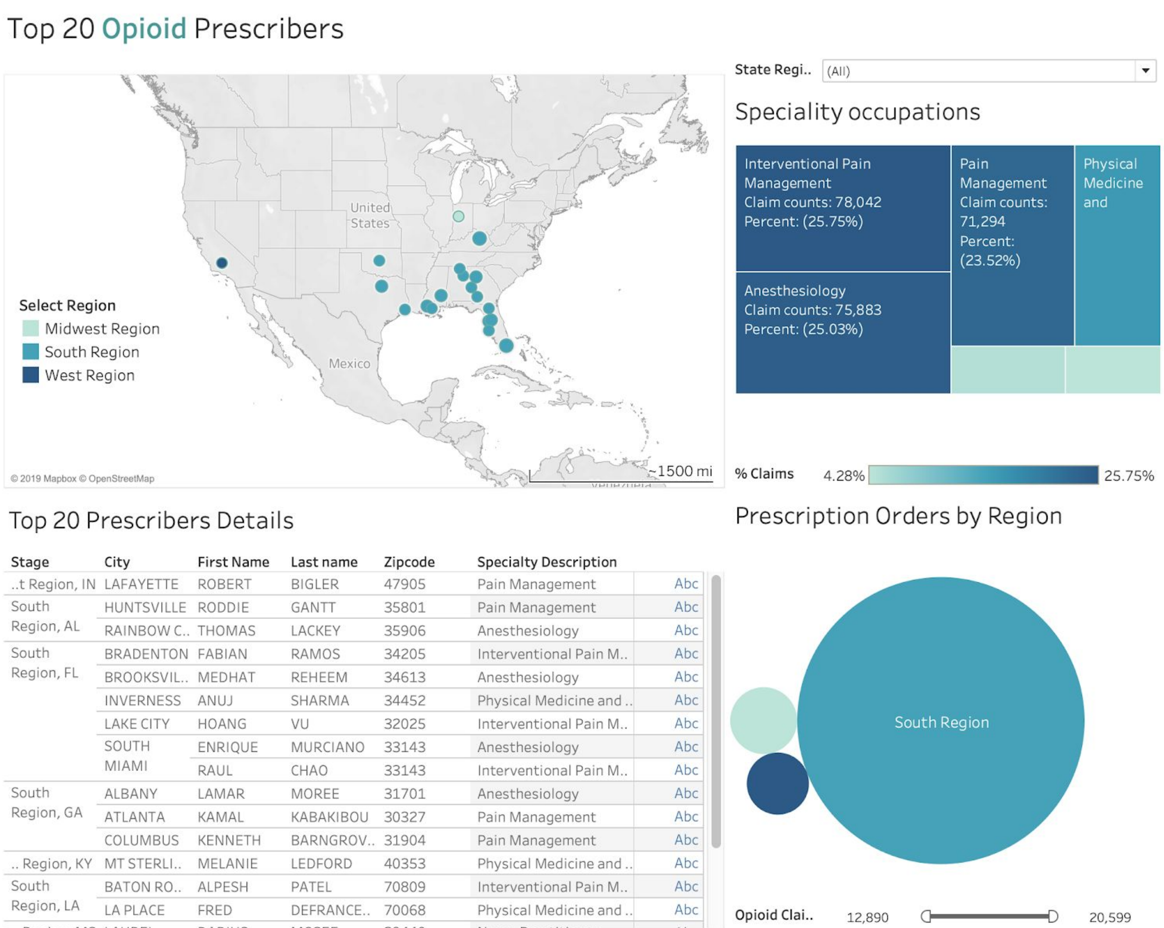


Global Residuals by Specialty: Does this hold for the outlier providers?

Outlier Visualization / Product

Goal of the visualization output is to allow end-users to get quick insights into which providers are outliers in their prescription patterns according to our models. The user will visually see who they are, where they are, combined with explanations from our model and clustering analysis. End state is an interactive dashboard that filters all related information across our key variables identified in our analysis. End users can explore all aspects of our analysis from original data to model predictions.

Final visualization product is still in development but several quick demos of this dashboard have been tested to determine which visualizations work best, how to store back-end data, and how to serve insights to end-users. Our strategy is to implement all preprocessing before piping our data into our visualization medium. Ultimately we are targeting compact, simple, intuitive plots to engage the end user.



Planned Activities

As per the proposal presentation we are on track and we worked on the activities as planned with some minor modifications

Proposed (old) plan of activities

Week	Activity
Wee 1-2	Develop a **standardized** training **dataset** based on provider data from the CMS API.
Week 3-4	Use **feature extraction** methods like vTreat to encode categorical data in a format suitable for predictive modeling.
Week 4-5	• Use various **clustering algorithms** to analyze the data. • Helps to **understand relationships** between provider attributes and prescription rates.
Week 5-6	**Develop a model** to estimate prescription rate based on provider attributes.
Week 6-9	• **Visualize** the outputs of predictive model. • Employ model interpretation methods like SHAP and LIME. • Enable identification of opioid PIP providers geographically using visualization toolkits such as D3.js, Plotly and ECharts.

Revised plan of activities

Week	Activity	Changes and Status
Wee 1-2	Develop a standardized training dataset based on provider data from the CMS API [5].	**Completed** by Data Team (Matt, Zach and Prashant)
Week 3-4	Use feature extraction methods like vTreat to encode categorical data in a format suitable for predictive modeling [6]. Addition - Visualization team - Learn and explore data using Tableau	**Complete** by Analytics team (Prashant, Supriya, Zach, Matt, Pallavi) Change - We used python libraries to transform and massage the data for modelling purposes.
Week 4-9	• Use various clustering algorithms to analyze the outliers data [12]. • Helps to understand relationships between provider attributes and prescription rates for outlier providers	**In Progress** by Analytics team (Prashant, Supriya, Zack, Matt, Pallavi) Clustering - In progress
Week 4-9	Develop a model to estimate prescription rate and opioid days supply based on provider attributes. Include SHAP [15] and LIME [16] methods and append data to original dataset for visualization team.	**Complete** (Matt, Zack, Prashant) Models training, SHAP values calculated, and sent to clustering and visualization team
Week 6-9	• Develop interactive data product to expose analysis to end users • Output will visualize original data, model, clustering and outlier analysis • Enable identification of opioid PIP providers geographically using tableau.	**In Progress** by Visualization Team (Supriya, Pallavi) Visualization team has drafted reports from outcomes of the predictive models

Conclusion and Discussion:

Great progress has been made towards delivering a cutting edge, easy-to-use, data-driven analytics project - which offers insights on opioid prescribing from complex, multi-dimensional data that covers over a million providers. By visualizing outliers and clusters, our project provides evidence, ranging from practices and trends that warrant further review, to possible provider over-scripting.

With our models complete, and clustering and visualization on the way, we are already seeing signs of possible real-world impact from our hard work. For example, one of the overscripting and overwriting outliers our model identified has already been found for medical malpractice in Las Vegas[26].

We are confident allowing end users to quickly explore and understand the complex nature of provider overscripting will help identify and stop more cases like this nationally, directly impacting the opioid epidemic.

Distribution of Effort: All team members have contributed similar amount of effort

References:

[0] Wilson, C. (n.d.). Drugs in America: Watch the Epidemic Spread. Retrieved from https://time.com/4260798/drug-epidemic-america/

[1] National Institute on Drug Abuse. (2019, January 22). Opioid Overdose Crisis. Retrieved from https://www.drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis

[2] Ellis1, R. J., Genes2, N., & Ma'ayan, A. (2019, January 29). Predicting opioid dependence from electronic health records with machine learning. Retrieved from https://biodatamining.biomedcentral.com/articles/10.1186/s13040-019-0193-0

[3] Schnell, M., & Currie, J. (2018). Addressing the Opioid Epidemic: Is There a Role for Physician Education? *American Journal of Health Economics*, *4*(3), 383–410. doi: 10.1162/ajhe_a_00113

[4] Matthews, D. C., Brillant, M., Jimoh, K. O., Singleton, W., McLean-Veysey, P., & Sketris, I. (2019). Patterns of opioid prescribing by dentists in a pediatric population: a retrospective observational study. CMAJ open, 7(3), E497–E503. doi:10.9778/cmajo.20190021

[5] Socrata Developer Portal. (n.d.). Retrieved from https://dev.socrata.com/foundry/data.cms.gov/6wg9-kwip

[6] Zumel, Nina, & John. (2019, September 22). vtreat: a data.frame Processor for Predictive Modeling. Retrieved from https://arxiv.org/abs/1611.09477

[7] Skala, K., Reichl, L., Ilias, W., Likar, R., Grogl-Aringer, G., Wallner, C., … Walter, H. (2013). Can we predict addiction to opioid analgesics? A possible tool to estimate the risk of opioid addiction in patients with pain. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/24284844

[8] P.GuyJr., G. (2019, February 16). Identifying opioid prescribing patterns for high-volume prescribers via cluster analysis. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0376871618306926

[9] (PDF) Hierarchical Latent Class Models for Cluster Analysis. (n.d.). Retrieved from https://www.researchgate.net/publication/228679557_Hierarchical_Latent_Class_Models_for_Cluster_Analysis

[10] A Fast Clustering Algorithm to Cluster Very Large ... (n.d.). Retrieved from https://grid.cs.gsu.edu/~wkim/index_files/papers/fastclusteringHuang.pdf

[11] Huang, Z. (1997, January 1). [PDF] Clustering Large Data Sets with Mixed Numeric and Categorical Values - Semantic Scholar. Retrieved from https://www.semanticscholar.org/paper/Clustering-Large-Data-Sets-with-Mixed-Numeric-and-Huang/d42bb5ad2d03be6d8fefa63d25d02c0711d19728

[12] Visualizing Data using t-SNE - cs.toronto.edu. (n.d.). Retrieved from http://www.cs.toronto.edu/~hinton/absps/tsne.pdf

[13] A Density-Based Algorithm for Discovering Clusters in ... (n.d.). Retrieved from https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

[14] Feature Selection with Ensembles, Artificial Variables ... (n.d.). Retrieved from https://www.researchgate.net/publication/220320233_Feature_Selection_with_Ensembles_Artificial_Variables_and_Redundancy_Elimination

[15] Lundberg, Scott, & Lee. (2017, November 25). A Unified Approach to Interpreting Model Predictions. Retrieved from https://arxiv.org/abs/1705.07874

[16] Ribeiro, Tulio, M., Singh, Sameer, & Carlos. (2016, August 9). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Retrieved from https://arxiv.org/abs/1602.04938

[17] Li, D., Mei, H., Shen, Y., Su, S., Zhang, W., Wang, J., … Chen, W. (2018). ECharts: A declarative framework for rapid construction of web-based visualization. Visual Informatics, 2(2), 136–146. doi: 10.1016/j.visinf.2018.04.011

[18] Slundberg. (n.d.). slundberg/shap. Retrieved from https://github.com/slundberg/shap/blob/master/README.md

[19] Story Map Series. (n.d.). Retrieved from https://cms-oeda.maps.arcgis.com/apps/MapSeries/index.html?appid=735f83ac6e984d6fade11b241d295585

[20] Understanding the Epidemic | Drug Overdose | CDC Injury Center. (n.d.). Retrieved from https://www.cdc.gov/drugoverdose/epidemic/index.html

[21] Bonnie, R. J., Ford, M. A., & Phillips, J. K. (2017). *Pain management and the opioid epidemic: balancing societal and individual benefits and risks of prescription opioid use: Committee on Pain Management and Regulatory Strategies to Address Prescription Opioid Abuse*. Washington, D.C.: The National Academies Press.

[22] Goodnough, A., Katz, J., & Sanger-katz, M. (2019, July 17). Drug Overdose Deaths Drop in U.S. for First Time Since 1990. Retrieved October 8, 2019, from https://www.nytimes.com/interactive/2019/07/17/upshot/drug-overdose-deaths-fall.html.

[23] J.Stopkaa, T., R.Kaplana, A., K.H.Chuia, K., Y.Walleybc, A., R.LaRochellec, M., & J.Rosecd, A. (2019, April 11). Opioid overdose deaths and potentially inappropriate opioid prescribing practices (PIP): A spatial epidemiological study. Retrieved from https://www.sciencedirect.com/science/article/pii/S0955395919300908

[24] Gregory L Stuart, Ryan C Shorey, Christopher R France, Jenny Macfie, Kathryn Bell, Kimberly B Fortner, Craig V Towers, Paul Schkolnik and Susan Ramsey.(2018, May 23). Empirical Studies Addressing the Opioid Epidemic: An Urgent Call for Research. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6090487/pdf/10.1177_1178221818784294.pdf

[25] Wei-Hsuan Lo-Ciganic, James L. Huang, Hao H. Zhang, ...Walid F. Gellad. (2019, March 22). Evaluation of Machine Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries with opioid Prescriptions. Retrieved from https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2728625

[26] Las Vegas Sun. (2012, May 18). Retrieved from https://lasvegassun.com/news/2012/may/18/doctor-involved-medical-malpractice-kickbacks-prob/