Data & Visual Analytics Final Report

Nopioid: Targeting High-Residual Opioid Prescription Providers

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

Project Objective

Authorities need to identify where resources should be targeted to reduce opioid prescribing. 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 are sufficient for 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], which ignores the role the provider takes in the crisis.

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].

Proposed Methods:

***Intuition and Innovation**:*

Existing public, government visualizations show *opioid prescription* 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:

- Advanced modeling, analysis and visualization of prescription rates, prescription length at the provider level
- Utilization of SHAPLEY 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
- 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, as well as zip-code data (population density, rural/urban classification).

Two targets were identified to model:

- Opioid Prescription Rate
- Opioid Days Supply

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 less predictions) will be our focus - these are the providers who are over-scripting. Visualization into our modeling and advanced clustering analysis on these outliers will be our main product deliverable, showing insight into who, where, and why these providers are selected as outliers.

Evaluation / Experiments

Key Questions and Answers based on our Data Product:

- Who are the high residual providers in terms of prescription rate and days supply?
    - Models identify over 9K high residual outliers across the US
- Where are these outliers located?
    - Highest concentration of outliers: Northest, Southeast , South West
- What are the key variables that drive our complex model's predictions?
    - Credentials, Specialty, Region, State, Beneficiary Age
- How does our model predict at the individual provider level?
    - Mean Absolute Error: Prescriber rate model = 1.5%, opioid day supply model = 2,160 Days
- In the group of high residual outliers - are the specific segments?
    - Five segments found clustering around credentials and specialty

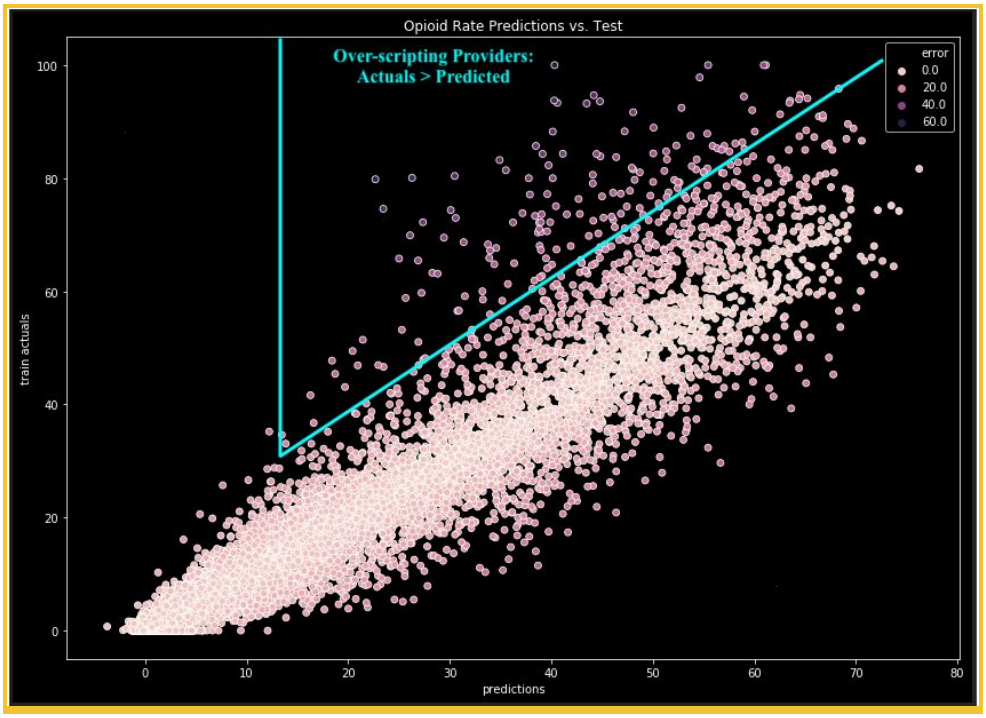
Details:

- Developed a standardized training dataset connecting together all medical practice attributes based on provider data from the CMS API. [5] Data was stored within an sqlite database, indexed and fragmented to allow for joining across the original dataset, prediction dataset, analysis dataset, and a combination of each - all to provide a lightweight dataset to visualize in our exploration interface.

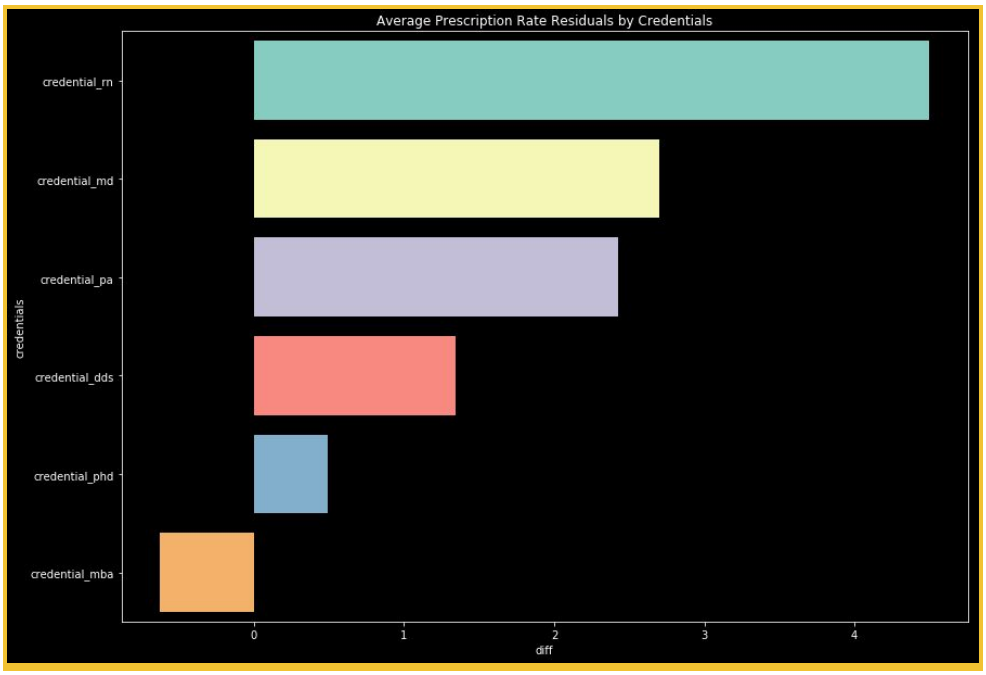
- Cleaned and encoded all 1 million plus provider records into a format that allows for predictive modeling and analysis. Careful preprocessing was completed to handle features with 50+ levels into a format that models could easily be iterated on. Methods such as one hot dummy variable encoding, variance filters, and recursive feature engineering were used to build the final training and analysis dataset. These methods allowed for us to reduce the feature-space from 500+ possible variables (high cardinal categorical features casted to dummy variables) to an information rich set of around 50 features.

- Modeled both provider opioid days supply and prescription rate using Gradient Boosting Trees Regression, complete with bayesian hyperparameter tuning to glean predictive extraordinary results. Designing an accurate model allows us to hone in on “true” high residual outliers in our data - those providers who are over-scripting, rather than residuals due to model skill.
    - Boosting methods grow tree based model iteratively, at each step re-weights the observations incorrectly labels from the last iteration.
    - All machine learning methods have parameters that are not native to the fitted model, these are hyperparameters.
    - Selecting the best values usually involves analyst intuition, heuristic based “starting points” or a brute force grid search - testing a huge range of possible values and then selecting the best based on an out of sample performance metric.
    - Bayesian Optimization will still iteratively try out a range of possible values but will “jump” to the next hyperparameter value based on expected probability of improvement. This is especially important for problems where our objective function without a closed form, is expensive to evaluate, and / or is noisy. Due to the high level of categorical values we believe all of these to be true for our dataset.
    - Results of a randomly samples validation set mean absolute error is estimated to be 1.5% for prescription rate and 2,160 days for prescription length.
    - All 1 million plus records were then scored using our models and residual values were calculated, resulting in our analysis dataset, the complete list of high residual outliers.
    - SHAPLEY value calculations for each prediction were appended to the dataset. SHAPLEY contributions allow for provider level inference into why the model is giving its individual prediction. Gains and costs of each prediction are spread across each feature in the feature-set. SHAPLEY interpretation goes a long way to demystify the results of our models traditionally thought of as “black boxes”.

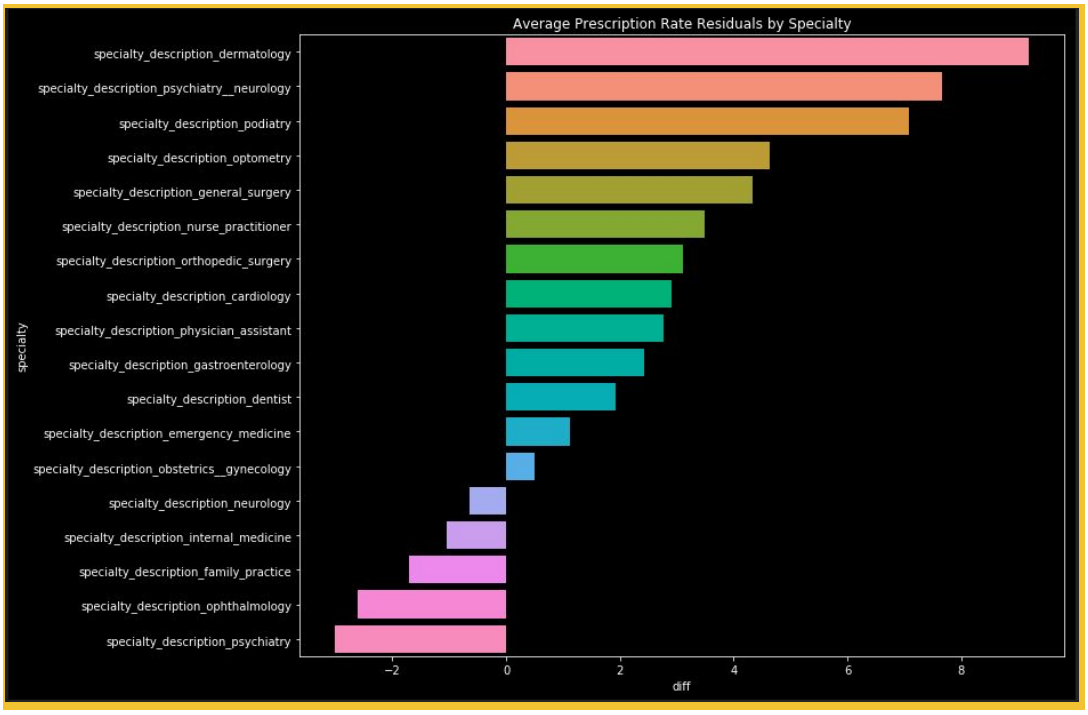
Prescription Rate Model Residuals: Showing top outliers of prescription rate by practice



Prescription Rate Residuals by Credentials: Registered Nurses (RN) have highest prescription rate residuals across all credentials



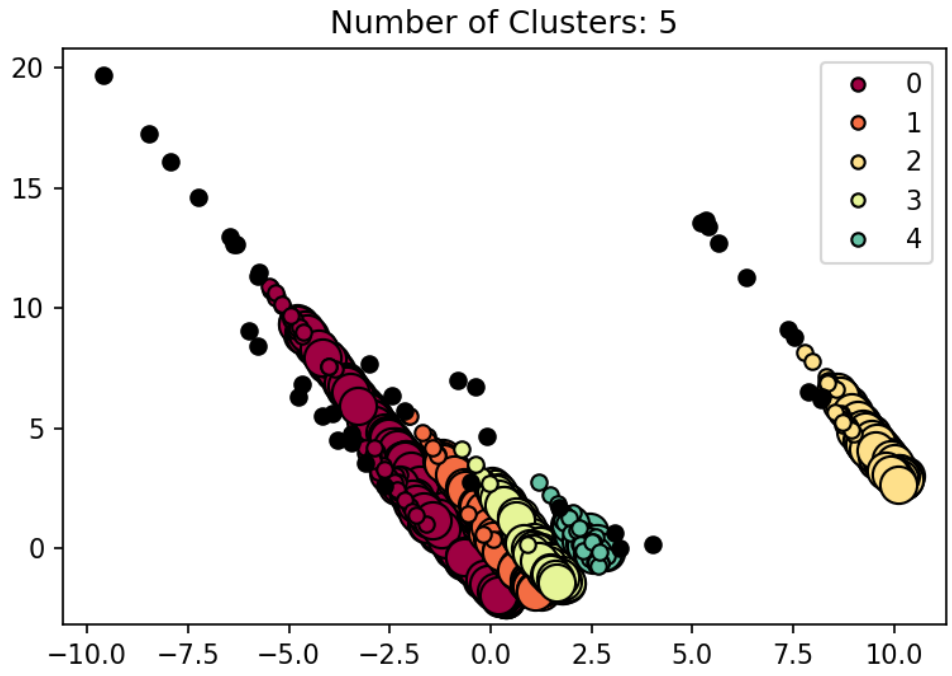
Prescription Rate Residuals by Specialty: Dermatology, Neurology among the highest residual prescription rate specialties



- State of the art clustering techniques, and well as tried and true methods of exploratory data analysis were performed on the outliers dataset.
    - **DBSCAN**, a density-based clustering algorithm, segmented the outliers dataset into 5 interpretable clusters based on key variables such as opioid

claim count, beneficiary features, days' supply, prescription rate, credentials, specialty, and region.

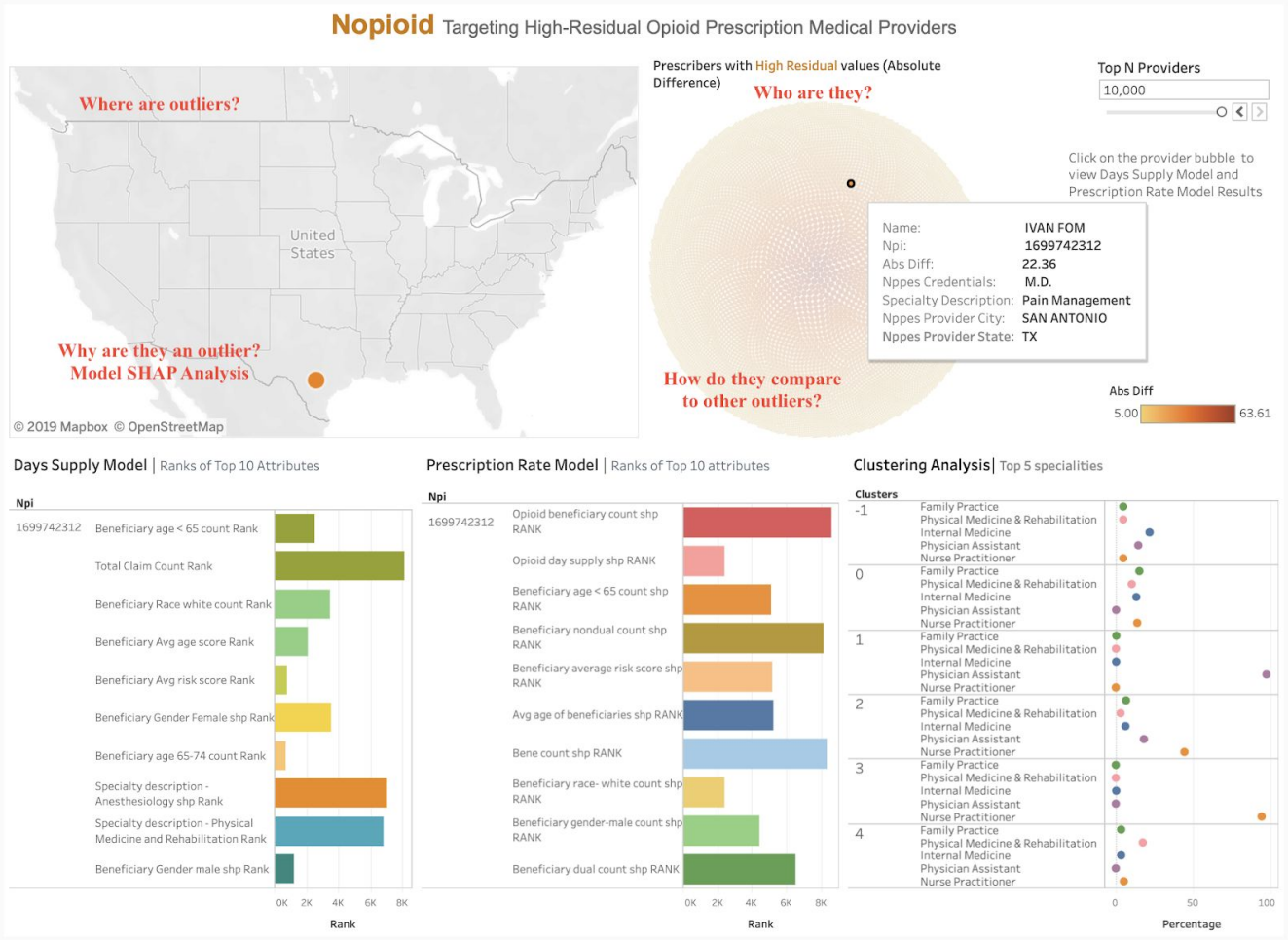
- Dimensionality reduction was performed through PCA (principal component analysis) and key components were selected which covered more than 90 % of the variances in feature space for the DBSCAN clustering model.
- There were four linear clusters and one compact cluster. A silhouette coefficient of 0.3 suggests that we have a fairly non-overlapping cluster with minimal outliers (41 black points).
- We observed that for all 5 clusters, the credential, specialty and region of the npi providers are the most influential features and influence significantly the clustering determination.



DBSCAN clusters: Based on credential, specialty, region, beneficiary characteristics and prescription rate

- An interactive insights dashboard was created building off our model and analysis results. The dashboard effectively visualizes the high residual outliers across multiple dimensions.
    - Dimensions: geography, prescription rate feature contribution, days supply feature contribution, DBSCAN clustering labels, cluster summary statistics, and unique meta-data for each provider.
    - All results pivot and filter interactively upon click ultimately showing:
        - Who and where are the outliers
        - Why do we think they are outliers (model inference)
        - How do they compare to like peers in the outlier dataset (clustering)
        - Results are succinctly presented for intuitive end-user consumption

Nopioid Dashboard allows end users to explore and answer their own key questions



[Nopioid Dashboard]

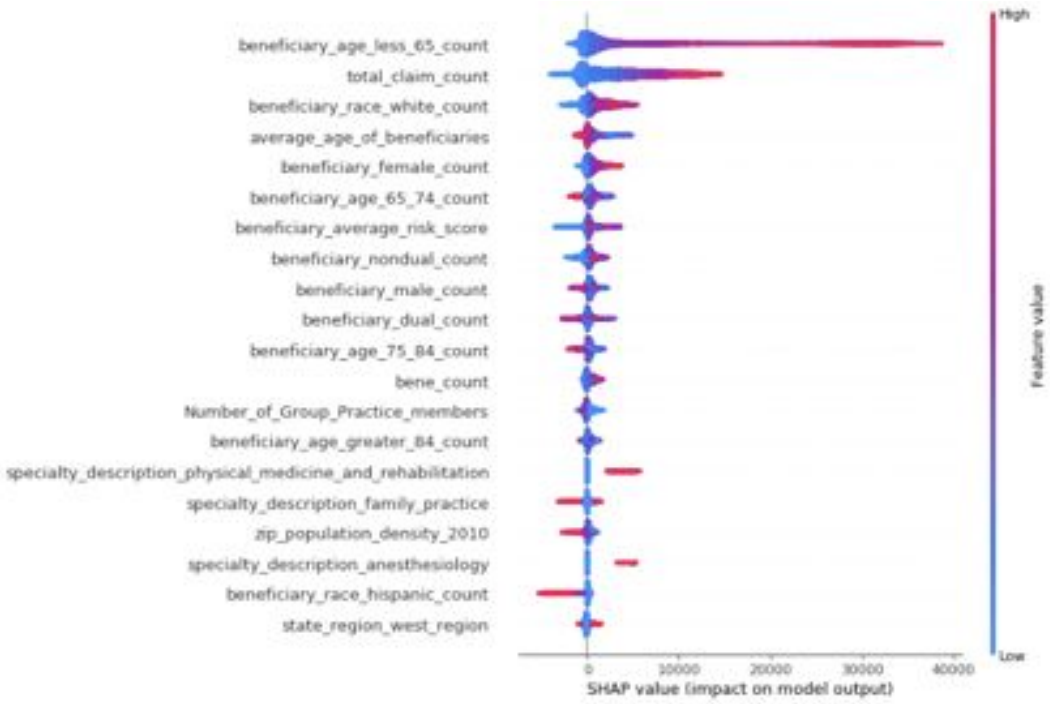
Deep Dive: Exploring the attribute "Beneficiaries Under the Age of 65”

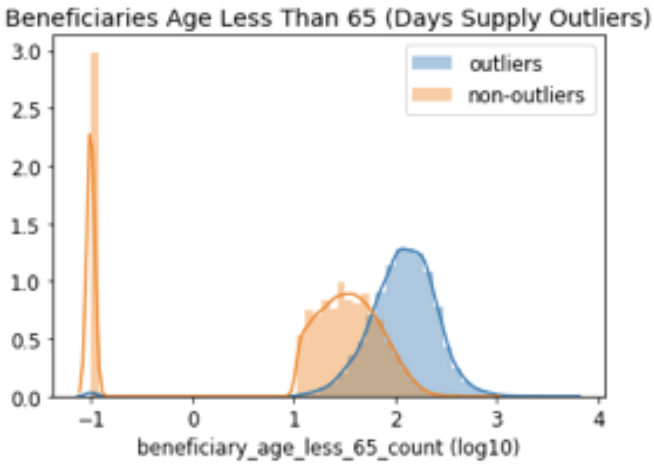
The most impactful attribute in the opioid prescription days supply model is the provider's number of beneficiaries under the age of 65. Interpreting the SHAP chart, we see a cluster of blue (low impact) points close to zero, and a tail of red points stretching to the right.

Another reason for exploring this attribute is that Medicare beneficiaries under the age of 65 represent a special population. In order to qualify for Medicare before age 65, the patient must have a disability.

After obtaining the model predictions for opioid days supply, we further define a provider as an outlier if the provider’s residual error for predicted opioid days supply is +/- 1 standard deviation from the mean error. We see that providers with large patient panels of beneficiaries under the age of 65 are represented higher in the group of outlier providers.

In the continuing discussion, we will look specifically at the group of outlier providers. Recall the earlier discovery that there is a cluster of providers that have high SHAP values for patients under 65. We find that about 65% of patients in this cluster are on opioids, compared to 44% for the smaller SHAP group.





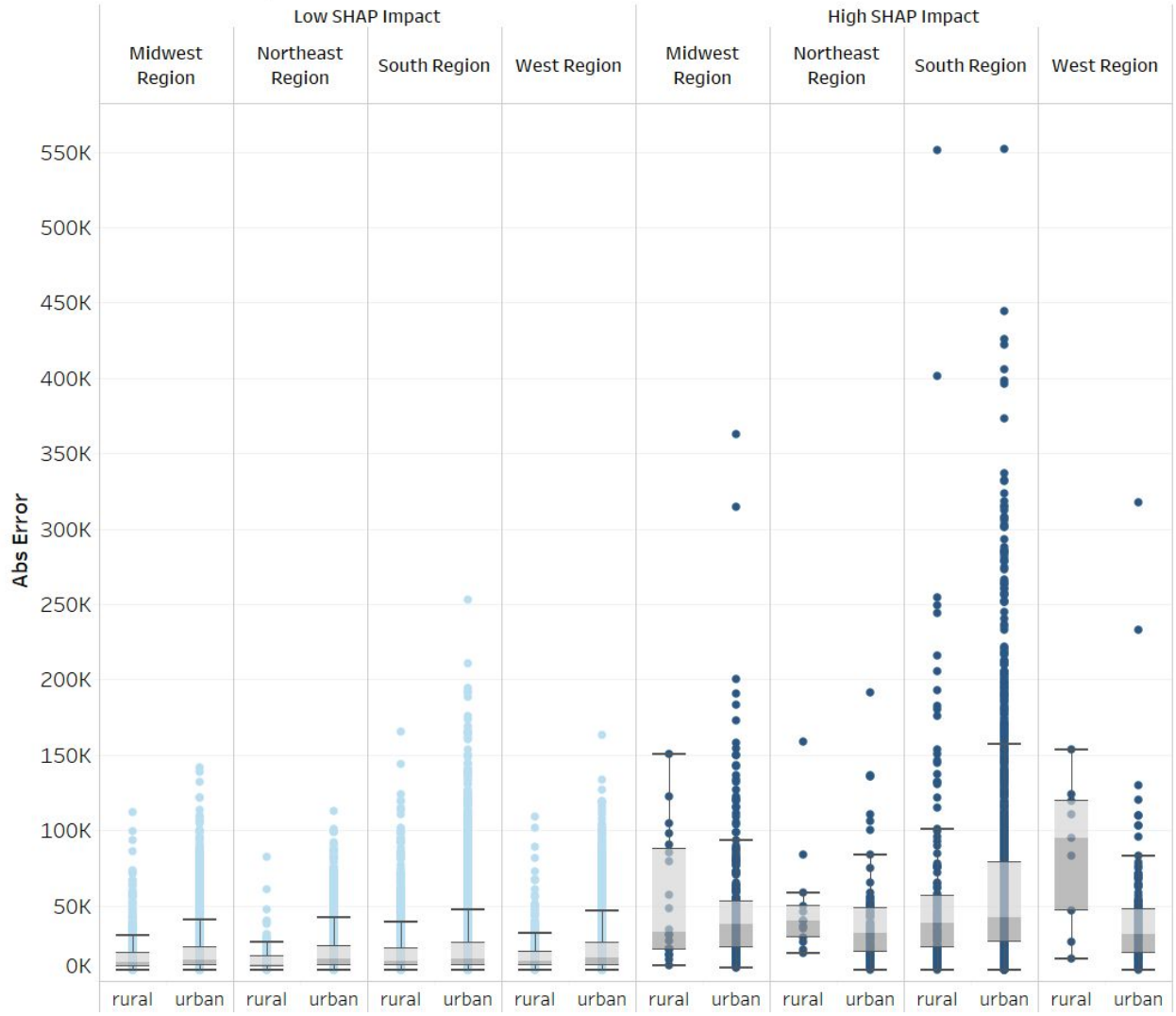
Given that this patient population is sicker, with disabilities and higher opioid use, a question may be whether there is variation by United States geographic region. Income and the number of providers will be more limited in rural areas. The following is a box plot analysis for the absolute residual error, by the following groups and attributes:

- Low SHAP Impact / High SHAP impact
- United States geographic region
- Urban / rural zip code

Each dot is a provider. We see that the boxplots overlap for most of the groupings, indicating that there does not appear to be differences between the regions. However, one boxplot does stand out – “West Region – Rural”. It stands out the “rural” boxplot does not overlap with the “urban” boxplot.

In exploring the outlier provider, the dot along the boxplot whiskers, it was found that the provider previously had his license suspended for 3 years related to lying under oath in malpractice claims. [24] This observation is suspicious but is not conclusive evidence that the provider is over prescribing opioids.

Abs Diff of Errors, by Low/High SHAP impact of Beneficiary Count Less than Age 65, Region



Conclusion

The end result of this project is an end-to-end data product that explicitly allows end-users to find, understand, and explore providers contributing to the opioid crisis. Our solution is cutting edge, easy-to-use and it a positive step towards solving the opioid epidemic.

The opioid insights dashboard offers insights on opioid prescribing from complex, multi-dimensional data that covers over a million providers. This data product will enable exploration of outliers, and build trust in our solutions by exposing the “why” of our model’s predictions.

Real-world validation of our solution is already evident. One of the possible over-scripting outliers our model identified had been previously found for medical malpractice in Las Vegas. We are confident allowing end users to quickly explore and understand the complex nature of provider over-scripting we 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

Reference List

[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] Before the Medical Board of California Department of Consumer Affairs State of California Retrieved from https://www2.mbc.ca.gov/BreezePDL/document.aspx?path=%5CDIDOCS%5C20140522%5CDMRAAAEC8%5C&did=AAAEC140522225005063.DID&licenseType=C&licenseNumber=50264

[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