Forecasting Medicaid Prescription Drug Demand

Capstone Project for BrainStation Data Science Diploma

Submitted to BrainStation
Submitted on 11 April 2022
Prepared by Albert King

Introduction

Drug shortages directly contribute to preventable deaths, drug stockpiling and rationing by hospitals and governments, and inflated drug costs. Studies[1] indicate that shortages arise due to myriad factors, including discontinuation of production, disincentivizing manufacturing costs, pipeline disruptions, and demand. Many of these factors are opaque or unpredictable, however, some data are available reflecting demand. United States Medicaid prescription drug data are available from 1991 into 2022, reporting number of prescriptions and reimbursements from the federal government to states. This project seeks to use these data to forecast demand for drugs in shortage using time series analysis models.

Research Objective

The objective of this project is to model and forecast demand for drugs susceptible to shortage based upon Medicaid prescription data in order to anticipate demand/need. Of relevance is how data analysis and machine learning techniques can generate these predictions. Reliable forecasts may help address shortages by anticipating demand, compared to straight line or mean predictions of current trends. In the future, the Medicaid analysis could be expanded to account for state-specific populations and Medicaid usage to further generalize the model.

Data Description

Data were collected from the Medicaid website (https://data.medicaid.gov/datasets?theme%5B0%5D=State+Drug+Utilization) as CSV files for each year 1991-2022. All data are reported by quarter. Each file contains the first 10 digits of a drug's generic name, its National Drug Code (NDC) number, the state and quarter corresponding to a given drug reimbursement, number of prescriptions, number of doses, and amount in dollars reimbursed. In this format drugs are duplicated across states and across quarters, resulting in 102,997,079 rows total. Some drugs are present without prescriptions in a given year. Further detail of the features and dataset shapes are discussed in Notebook 1. Medicaid notes that data are updated annually, though 2022 data were incomplete when downloaded (quarter 3 is missing data for most states, and quarter 4 is not present).

The US FDA publishes a list of recent, active, and pending drug shortages, updated daily, on their website. The data used in this project's final analysis were downloaded as a CSV file from https://www.accessdata.fda.gov/scripts/drugshortages/default.cfm on 07 April 2023 at 20:47. The relevant feature of this dataset is generic drug names. Status of the shortage (resolved, active, or pending), date of its most recent update, NDC numbers, manufacturer, and therapeutic application are also included in the dataset. Of 392 unique drugs listed as in shortage, 336 are shared between the Medicaid and shortage lists.

In the future, the project may be expanded to correlate drugs by NDC numbers and aggregate them by therapeutic application. The NDC number for a drug defines the substance, manufacturer, and dose; as such, any given drug relates back to many NDC numbers. There are significant discrepancies over time and across related formulations. Attempt was made to use data scraped from rxlist.com and downloaded from ncqa.org to match NDCs and therapeutic application, though this was minimally successful.

Methodology

Due to the size of the files, the Medicaid datasets were imported into a MySQL database either through command line or directly accessing the MySQL Server via SQLAlchemy in Python. During EDA, Tennessee, Washington, and South Dakota were found to have extremely high prescription reports in 2006 (TN and WA) and 2007 (SD). Data were removed from the database for these states in these years as outliers, confirmed by a Tukey's fence test for each state. Though this test found some additional values as "outliers", only the three visually apparent outliers (i.e. 4.8 billion prescriptions from WA in 2006) were removed to retain data. Further, data from 2022 quarters 2 and 3 were not reported for all states. Accordingly, I chose to drop Q2 and Q3 data from 2022 to ease modeling and maintain data integrity. No other changes were made to the database.

Data were interpolated using forward filling for states where there were no reported prescriptions for modeling federal level prescriptions. In the individual drug data, a line is used to determine if any points are greater than 5 times the overall median, as data are being batch processed and Tukey's test was observed to be very sensitive to the data. Any blanks in the data are then interpolated by forward filling. This may be revisited in the future to employ more rigorous statistical methods.

Regular expression comparison was used to relate drugs names in the Medicaid dataset to the current list of drugs in shortage. Because only 10 letters of the drug name are supplied from the Medicaid dataset, it would be preferable to relate them through NDC codes. Due to the difficulties discussed above, this method will augment matching in a future version of the project.

Parameters for the ARIMA function were determined experimentally by iterating through a series of potential values. Because some drugs include years where there are zero prescriptions (for example, if the drug was temporarily not available through Medicaid,) Symmetric Mean Absolute Percentage Error (SMAPE) was generally used to evaluate models as it is more robust to zero values compared to Mean Absolute Percentage Error (MAPE). For bulk processing, the ARIMA parameters corresponding to the 5 lowest SMAPE values (averaged between train and test prediction evaluations) were used to fit the data and the mean fit was used for the final SMAPE evaluation. For forecasting, future dates were fit with each set of parameters, and the mean of the 5 forecasts reported.

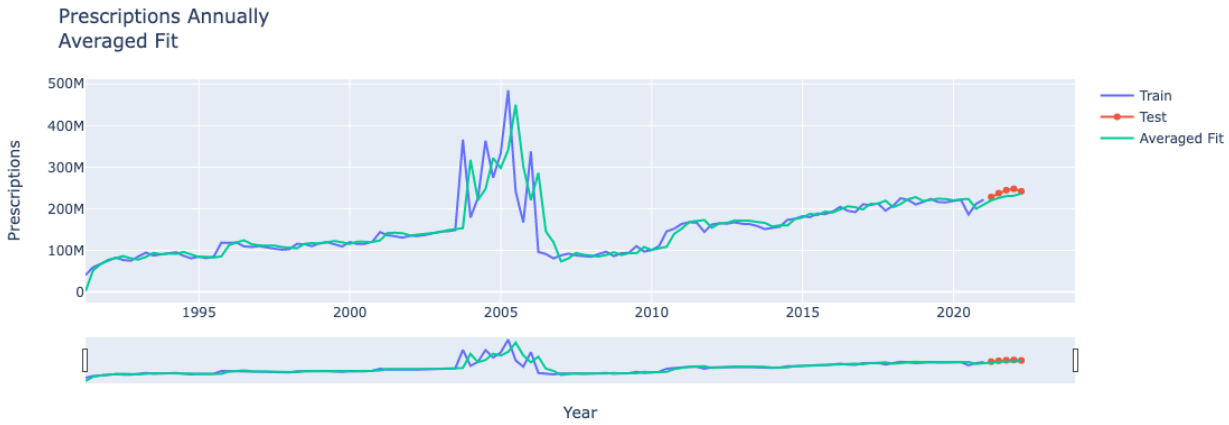
While ARIMA modeling is the focus of the project, other models are included for completeness. The Facebook Prophet model is optimized by iterating through the `changepoint_prior_scale` parameter, adjusting the number of inflection points available to the model. The RNN is essentially unchanged from the TensorFlow notebook model, simply optimized by adjusting the number of epochs.

Results

All three models produce relatively good evaluation metrics for test datasets, indicating reliability in their predictions forecasting approximately one year out. Using a constant prediction of the all-time mean of 153,236,124 per quarter for all federal prescriptions returns an SMAPE of 17.17%. This will be the baseline for other models to meet. The most optimized single ARIMA model returns a test SMAPE of 2.37%. Averaging the top 5 predictions results in a test SMAPE of 2.64%. This means that the ARIMA model fits the data better than a flat-line prediction. While the averaged model fits the training data significantly better than the top individual model, the test data increases slightly from 2.37% to 2.64% SMAPE.

While the Prophet model is not maximally optimized, adjusting "flexibility" in number of inflection points through the `changepoint_prior_scale` parameter allows some tuning of the model. The optimum test SMAPE was found to be 1.33%, a notable improvement over the ARIMA model and extreme improvement over the flat-line projection.

The RNN model does not currently provide a SMAPE, but does return Mean Average Error, MAE. In these terms the MAE is 0.1530 for the test dataset, reflecting a good fit at 1000 epochs.



Total quarterly prescriptions nationwide are fit using averaged optimized parameters above.

Conclusions

In conclusion, all models investigated are superior to straight-line mean projections for prescription drug demand. The recurrent neural network model deserves more attention as the initial model fits are highly promising. The ARIMA model is highly flexible and amenable to machine leaning styled hyperparameter optimization. While there are many aspects to optimize in future iterations of the project, correlating Medicaid forecasts to nationwide and state populations will help further address the concern of accurately forecasting drug demand to minimize shortages. Further, it would be deeply fascinating to utilize machine learning classification methods to investigate the relationship between demand and therapeutic application to learn how application may be related to shortages.

A total of 131 drugs present in the Medicaid data and the shortage list have been forecast to the third quarter of 2023. The modeling will be improved in the future to be more robust in dealing with data where the drug is new to the Medicaid data set, and a full history is not available. A slight downward trend is expected in prescriptions overall, from 95.4 million in the final quarter of 2022 to 93.9 million in the third quarter of 2023. The SMAPE is shown for each prediction, and the mean SMAPE for all drugs is 6.4%. The lowest SMAPE is 0.03%, and 74% of forecasts have a SMAPE less than or equal to 4.8%.

Reference

[1]Drug Shortages: Root Causes and Potential Solutions. A Report by the Drug Shortages Task Force. U.S. Food and Drug Administration. Published 2019. Accessed 22 February 20203 19:28 from https://www.fda.gov/media/131130/download