Predicting Drug Prices

Dana Hackel, DSI Capstone

TABLE OF CONTENTS

01

Key Terms and Abbreviations

Key Terms and Abbreviations

- Pharmaceutical product/ drug - substance with physiological effect
- National Drug Code (NDC) - unique 10 digit number assigned to a drug
- Over the Counter (OTC) - non-prescription
- Prescription (Rx) - requires written prescription from a doctor
- Dose - the defined amount of a drug to take at a given time
- Route - how the drug is taken
- Generic name - not a specific brand
- Trade name - specific brand name

02

Problem Statement

Problem Statement

The price of many drugs, both prescription (Rx) and over the counter (OTC) can often be very expensive. Having a way to predict price can help pharmaceutical companies determine price for newly approved drugs. It can also help companies compare their current price to similar drugs to determine if they are priced competitively. I will use the National Average Drug Acquisition cost data to build a regression model that predicts drug price values. I will also create an interactive graphic to help customers explore drug price factors.

03

Data

Data Sources

- NADAC - Center for Medicare and Medicaid Services
  - Updated weekly
  - November 2013 - present
- Orange Book -US Food and Drug Administration (FDA)
  - Approved drug products with therapeutic equivalences

Data Cleaning

- Outlier
- Removing repeat entries
  - Different “as of” dates
- Match both data frames
  - Fuzzywuzzy library
  - Levenshtein Distance - minimum number of single-character edits (insertions, deletions or substitutions)

Data Exploration

- Price distribution -
  - $1.53 average
  - $12.13 standard deviation
  - $0.0007 - $346.90 per unit
- Average price/ route
- Ingredient

Average OTC Drug Price by Route

Route	
NASAL	0.8557
ORAL	0.1610
RECTAL	0.0137
TOPICAL	0.0988

Average RX Drug Price by Route

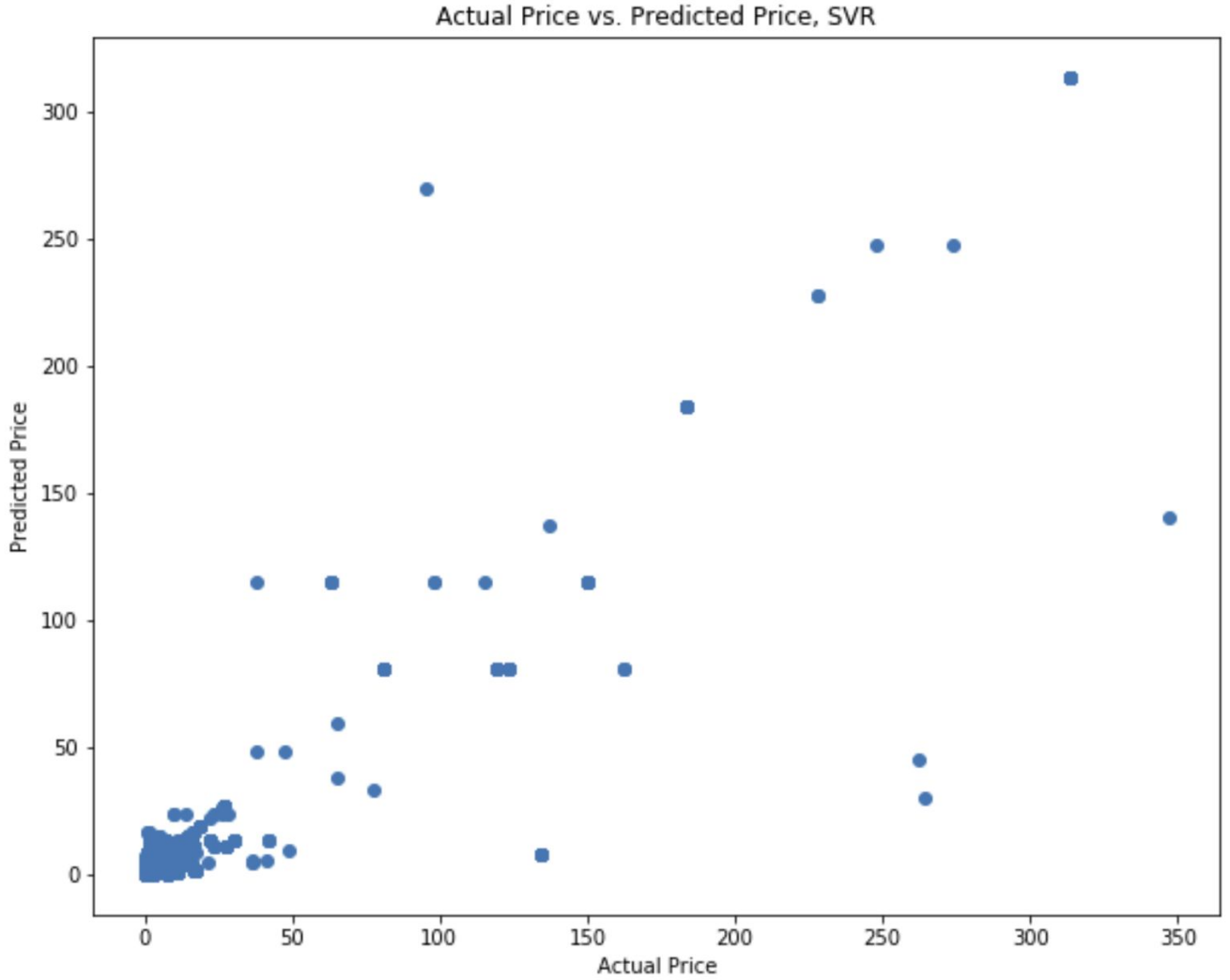
Route	
INHALATION	95.9
IRRIGATION	0.0
NASAL	118.3
OPHTHALMIC	22.7
ORAL	1.6
RECTAL	0.1
SPINAL	0.1
TOPICAL	0.8
VAGINAL	11.6

04

Predictive Modeling

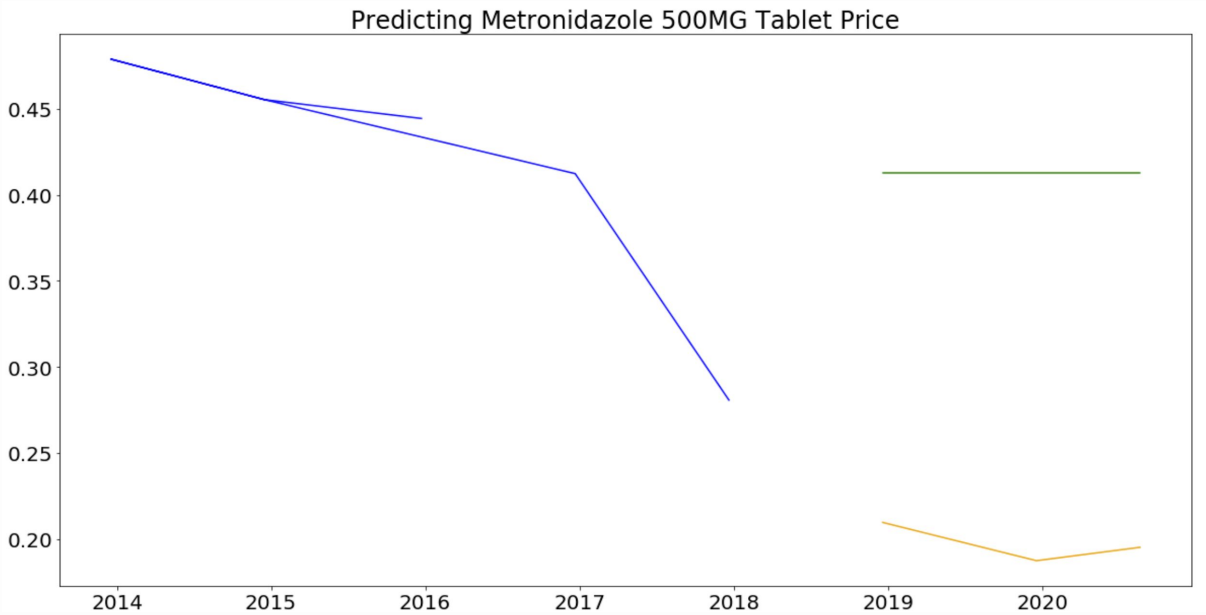
Price Prediction Results

Model	Test R2	Variation	RMSE
K Nearest Neighbor*	89.3%	3%	$4.25
Random Forest	79.8%	8%	$6.09
AdaBoost	64.6%	3%	$8.07
Support Vector Regressor*	89.3%	2%	$4.25


Actual Price vs. Predicted Price, SVR
Predicted Price
Actual Price
0
50
100
150
200
250
300
350

Price Prediction Over Time

- Many data points per date entered
- With averaged price - only 8 dates with information



05

Interactive Data Visualization

06

Conclusion and Next Steps

Conclusion

- Predict with ~89% R2
  - RMSEs still high

Next Steps

Immediate Actions:

- More drug prices
  - Updated prices
- Reduce variation and RMSE

Near Future:

- Refine data even more - timeseries
- Predicting ingredient - cluster categorical predictions

Other Fun Projects:

- Fuzzywuzzy and DNA matching

Sources

- https://data.medicaid.gov/Drug-Pricing-and-Payment/NADAC-National-Average-Drug-Acquisition-Cost-/a4y5-998d
- https://www.fda.gov/drugs/drug-approvals-and-databases/orange-book-data-files
- https://github.com/seatgeek/fuzzywuzzy
- https://github.com/alofgran/Drug-Price-Prediction
- General Assembly DSI Lessons
- Tableau Public

Thank you!

Feel free to contact:
hackel.dana@gmail.com

Github: https://github.com/DanaHackel

Twitter: https://twitter.com/Dana_Hackel

Medium: https://medium.com/@hackel.dana

Linkedin: https://linkedin,com/dana-hackel