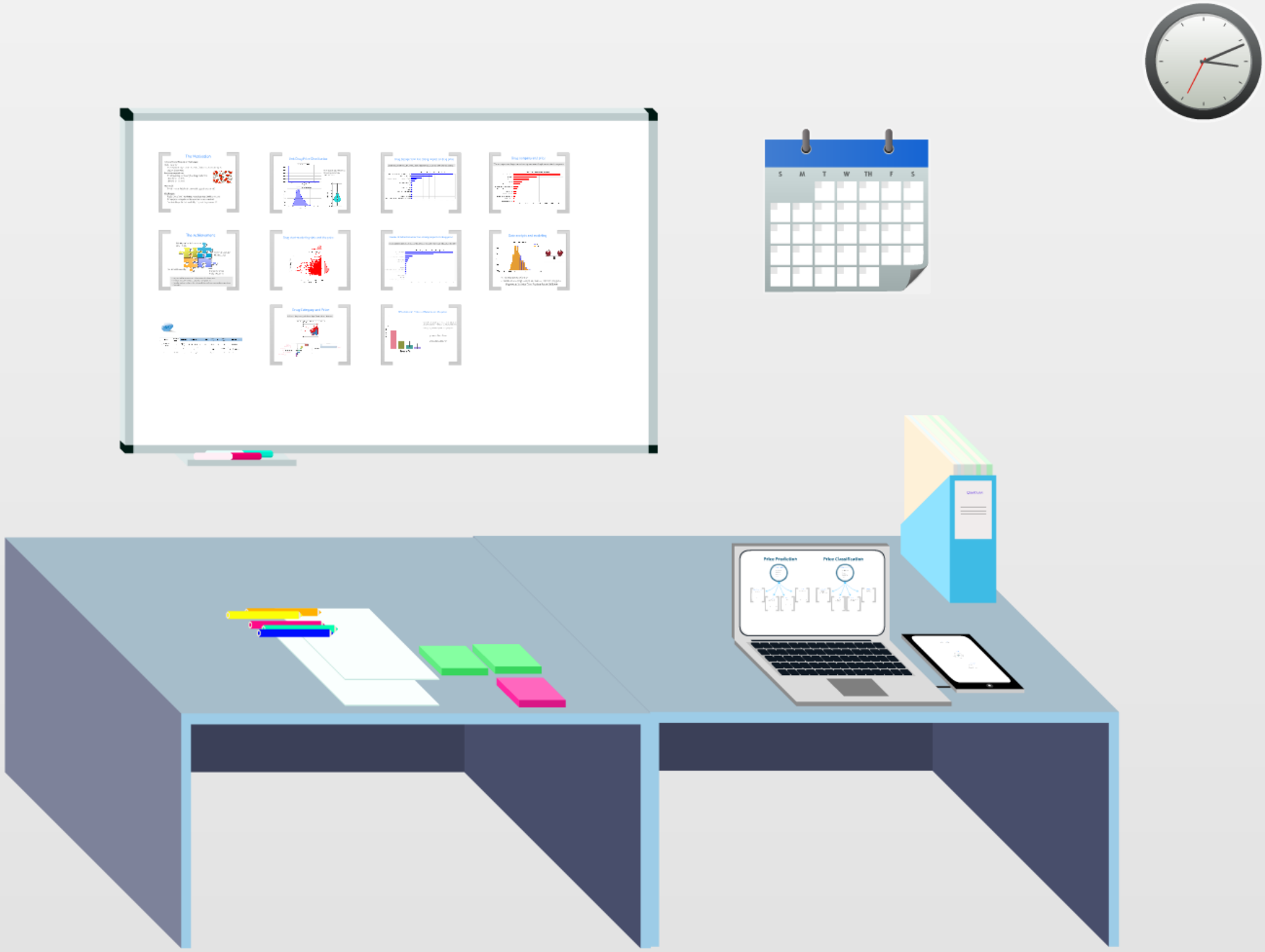
Pharmaceutics Meets Machine Learning: Predicting Drug Price by Simplified Approaches

Bo Lian, Choutine Zhou, Robin Gane-McCalla

Pharmaceutics Meets Machine Learning: Predicting Drug Price by Simplified Approaches

Bo Lian, Choutine Zhou, Robin Gane-McCalla


S M T W TH F S
Price Prediction
Price Classification

The Motivation

Life and Social Economic Challenges:

High drug price

Prescription drug prices in the United States have been among the highest in the world.

Enormous expenditure

Total spending on prescription drugs in the U.S

$425 billion in 2015

$600 billion in 2021

Our Goal:

Develop a simplified and universal drug prediction model

Challenges:

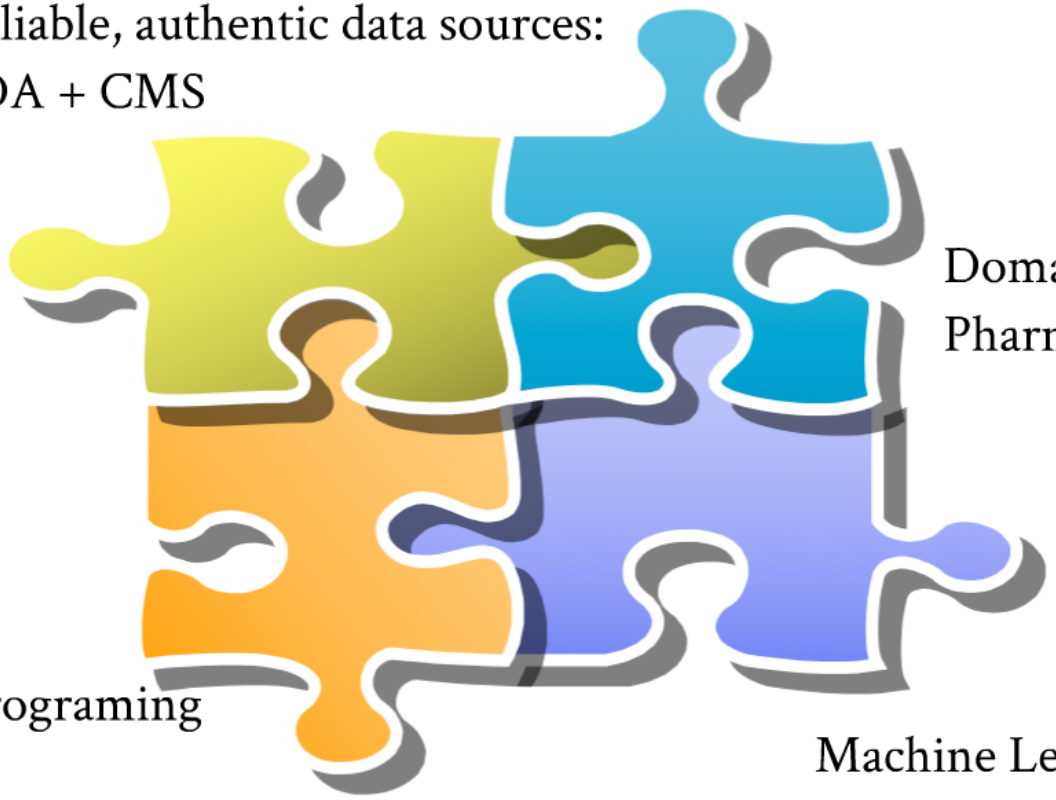
R&D, clinical trial, marketing, manufacturing, distribution cost

Policy/politics/regulation/business/economics involved

No desirable public data available, they cost huge amount!!!

The Achievement

Reliable, authentic data sources:
FDA + CMS



Domain Knowledge:
Pharmaceutics

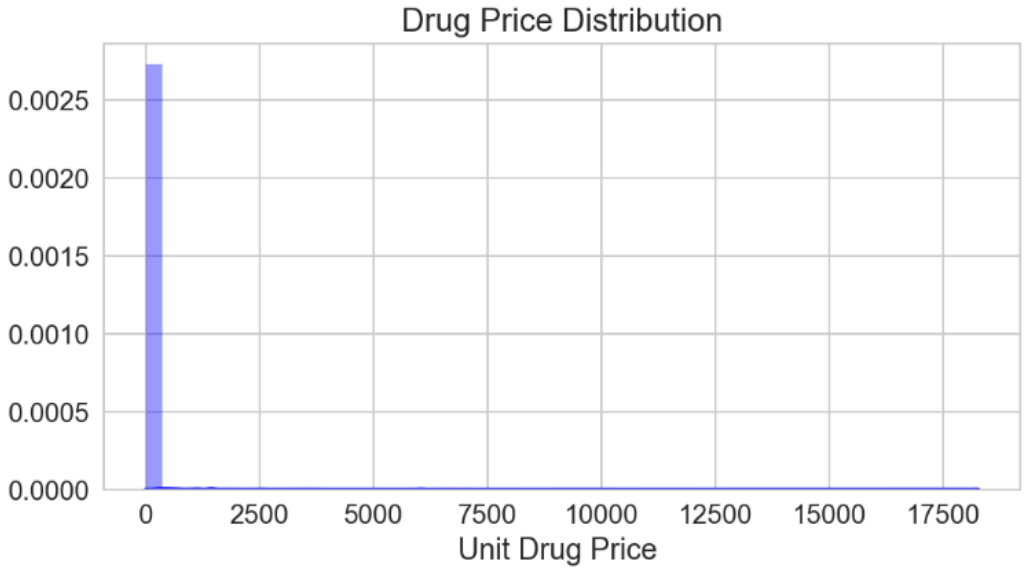
Statistics +Programing

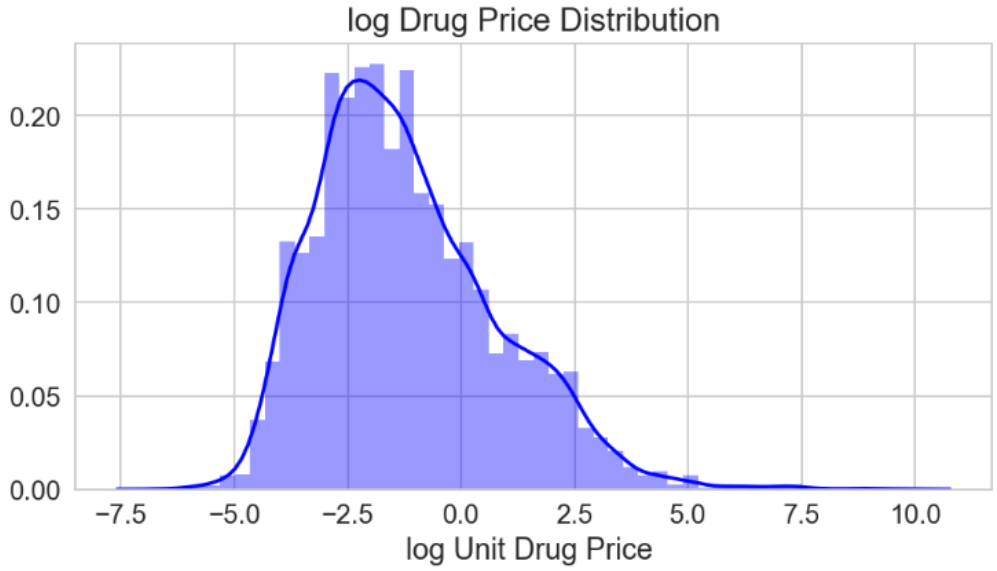
Machine Learning
Model Evaluation

- Fully utilized limited data info and imbalance data distribution
- 20+ Features scaled down to 7 , mostly are drug features
- 4 models built and evaluated for price prediction and price classification, respectively
- Innovation

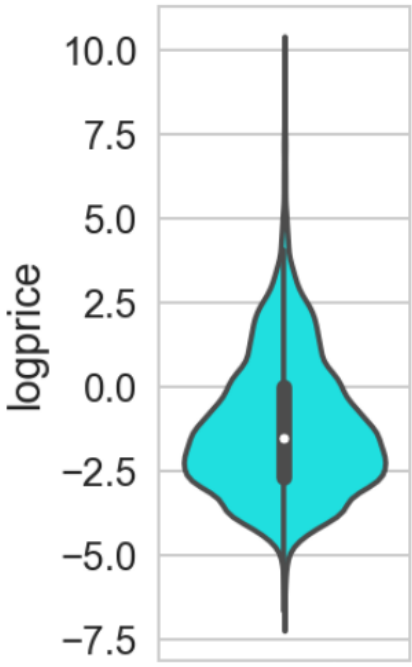
Product	Unit Price (dollar)	Company	Drug Category	Route	Dosage Form	DEA	Starting Date	Indication
Good neighbor pharmacy day time	0.13	Amerisource Bergen	Over-the-counter (OTC)	oral	capsule	N/A	8/2/12	cold and flu
Viagra	59.1	Pfizer	Brand chemical drugs (NDA)	oral	film coated tablet	N/A	3/27/98	Pulmonary hypertension Sexual dysfunction
Neulasta	9070.6	Amgen	Biologics (BLA)	subcutaneous	Injection	N/A	4/1/02	Cancer chemo-therapy

Unit Drug Price Distribution

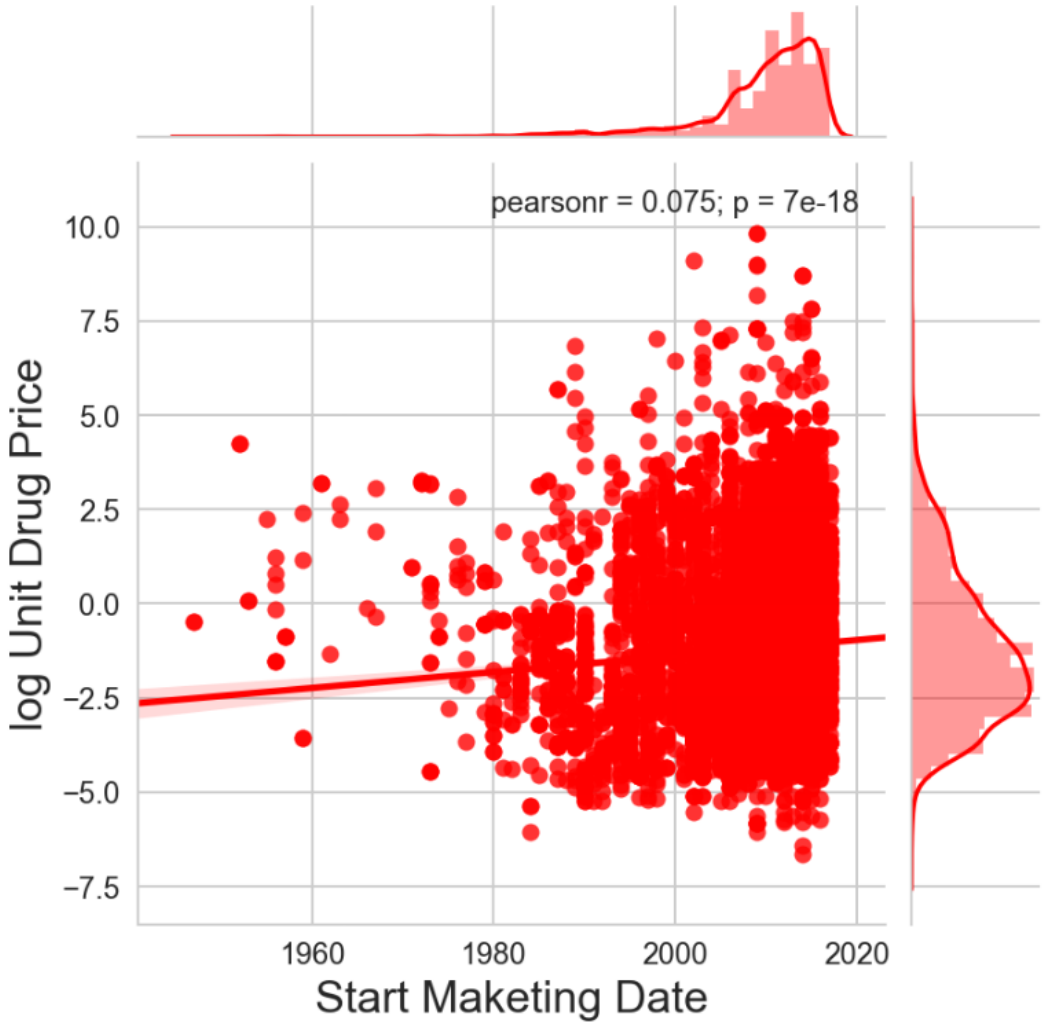




Right Skewed, Super Imbalanced
Most are low price drugs
Log transformed

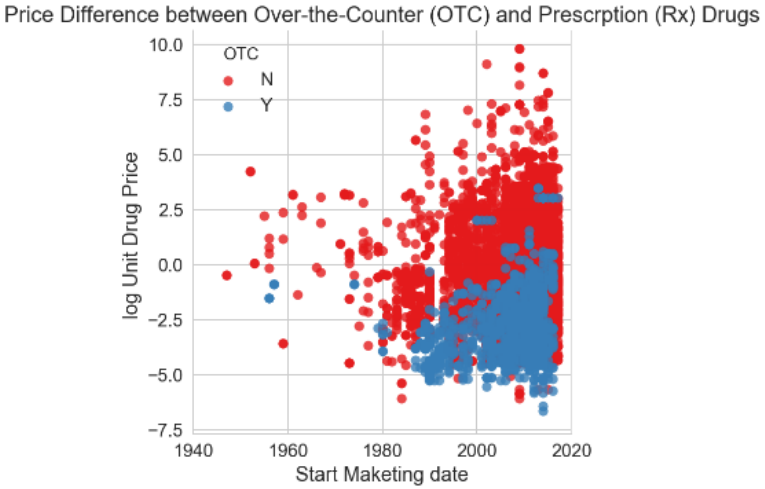


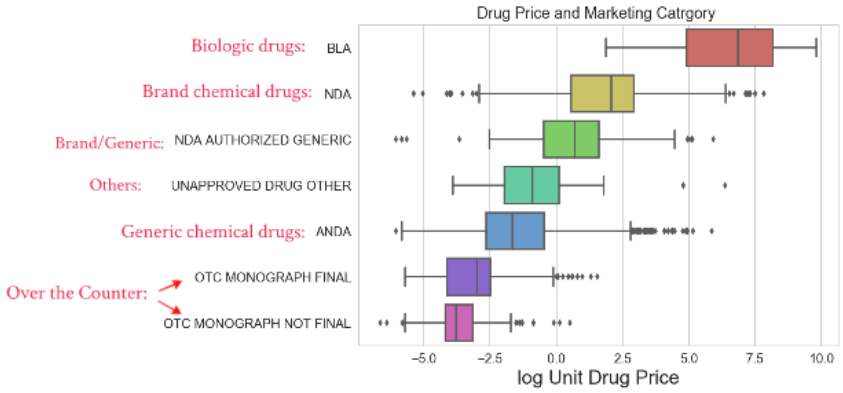
Drug start-marketing-date and the price

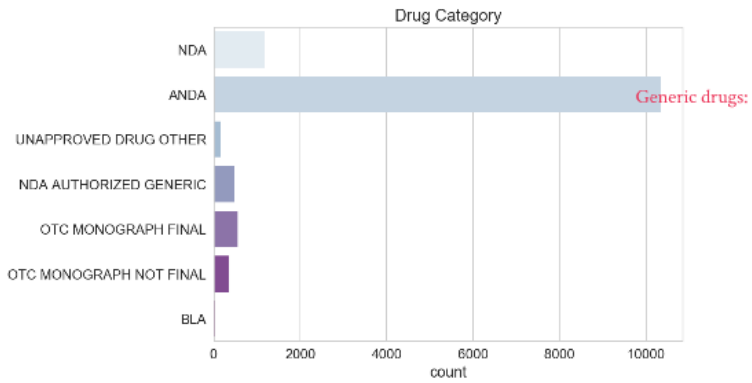


Drug Category and Price

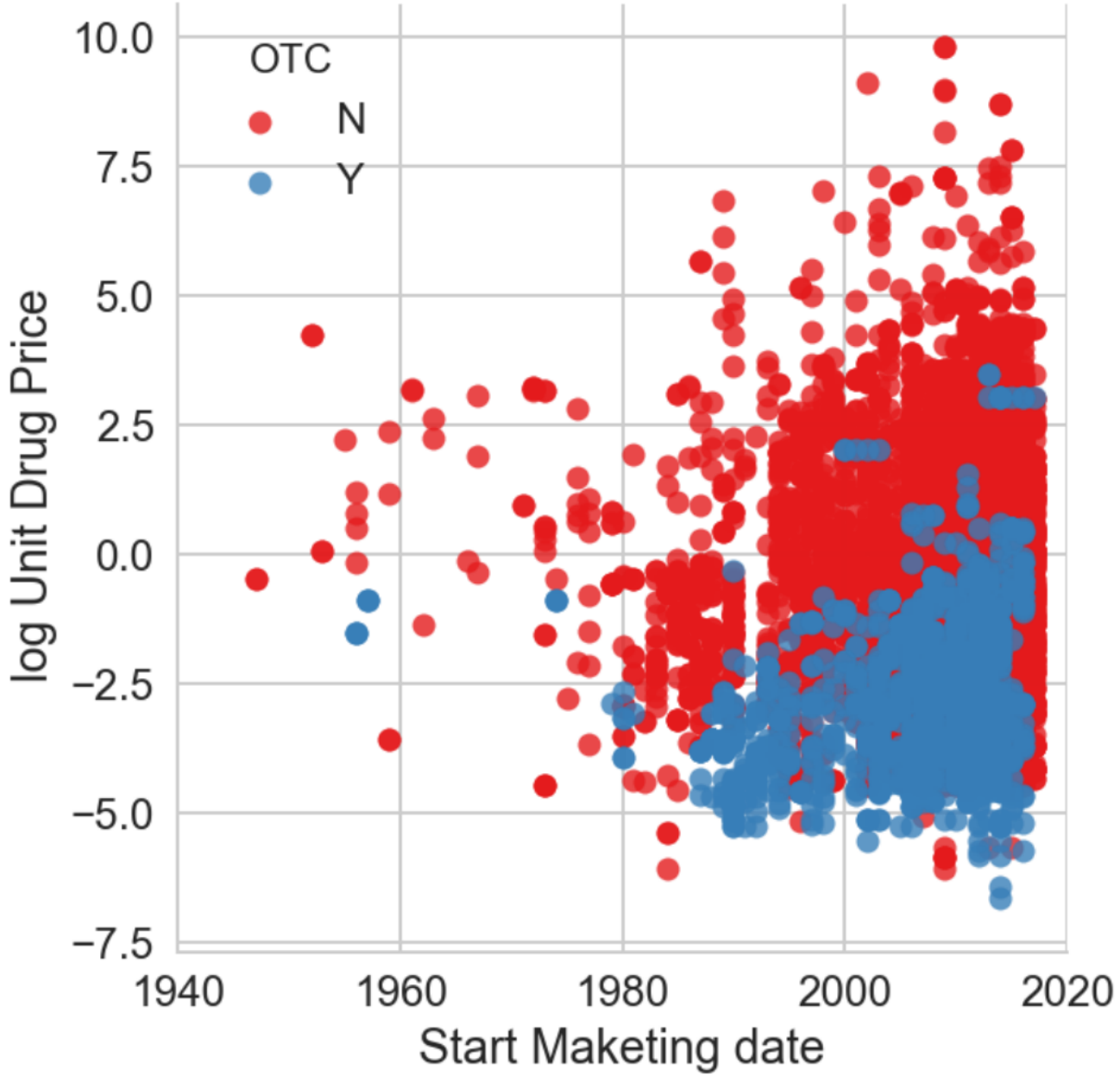
Different drug categories have significant price difference

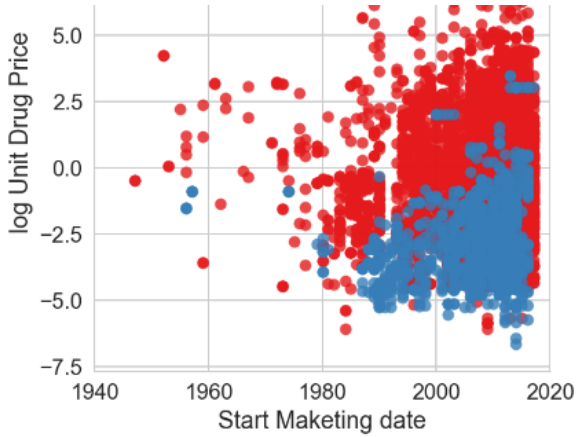


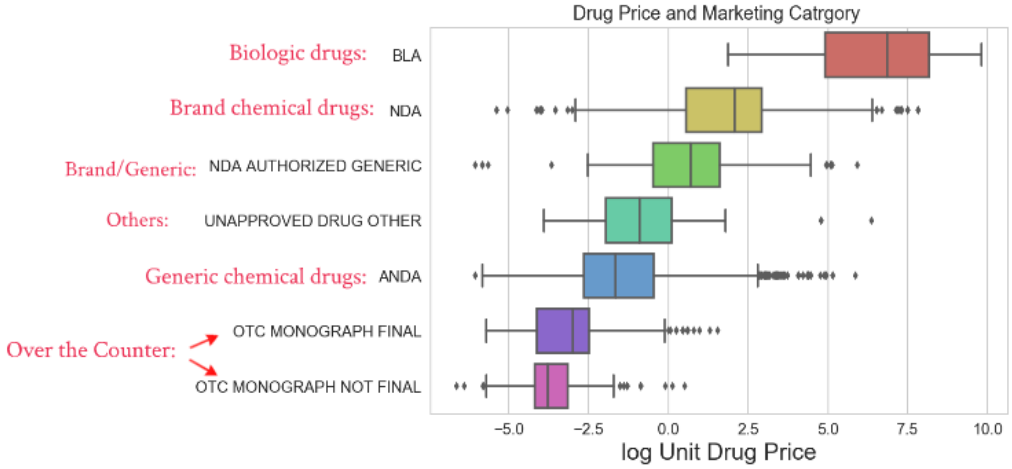


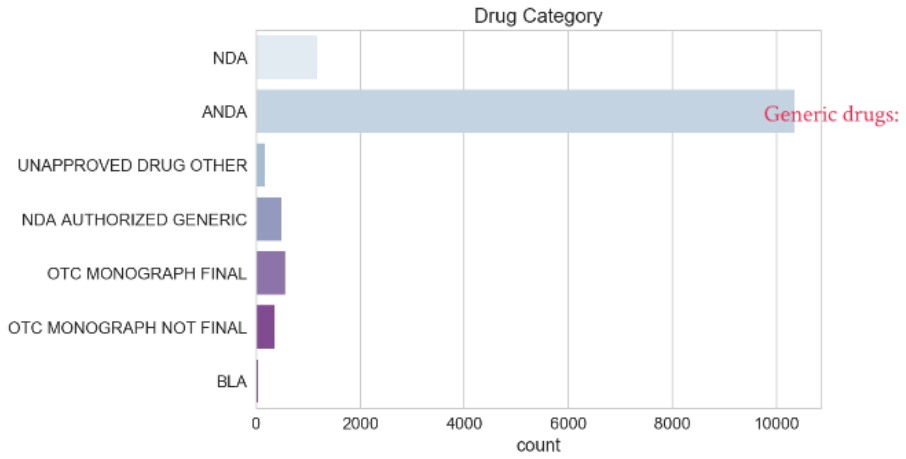


Price Difference between Over-the-Counter (OTC) and Prescrption (Rx) Drugs



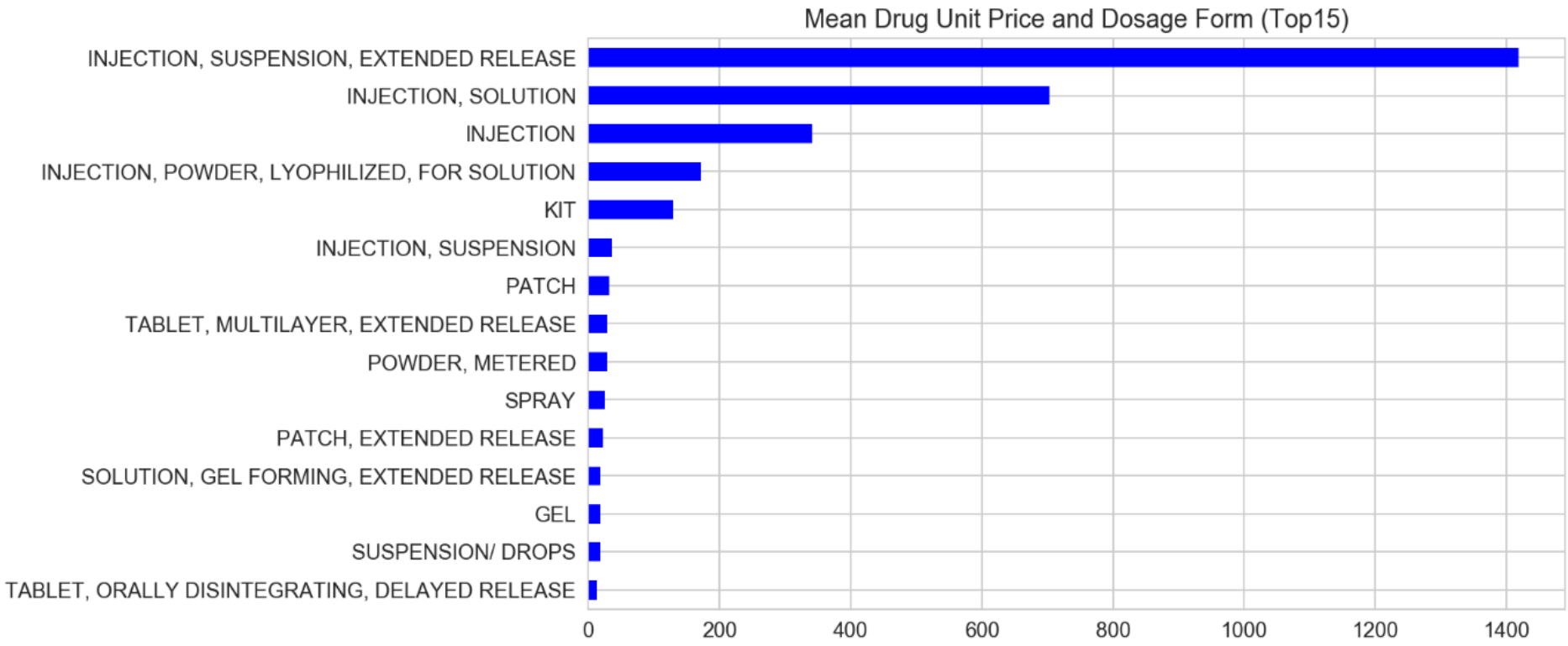

log Unit Drug Price
5.0
2.5
0.0
−2.5
−5.0
−7.5
1940
1960
1980
2000
2020
Start Maketing date


Drug Price and Marketing Catrgory
Biologic drugs:
BLA
Brand chemical drugs:
NDA
Brand/Generic:
NDA AUTHORIZED GENERIC
Others:
UNAPPROVED DRUG OTHER
Generic chemical drugs:
ANDA
Over the Counter:
OTC MONOGRAPH FINAL
OTC MONOGRAPH NOT FINAL
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
log Unit Drug Price


Drug Category
NDA
ANDA
Generic drugs:
UNAPPROVED DRUG OTHER
NDA AUTHORIZED GENERIC
OTC MONOGRAPH FINAL
OTC MONOGRAPH NOT FINAL
BLA
0
2000
4000
6000
8000
10000
count

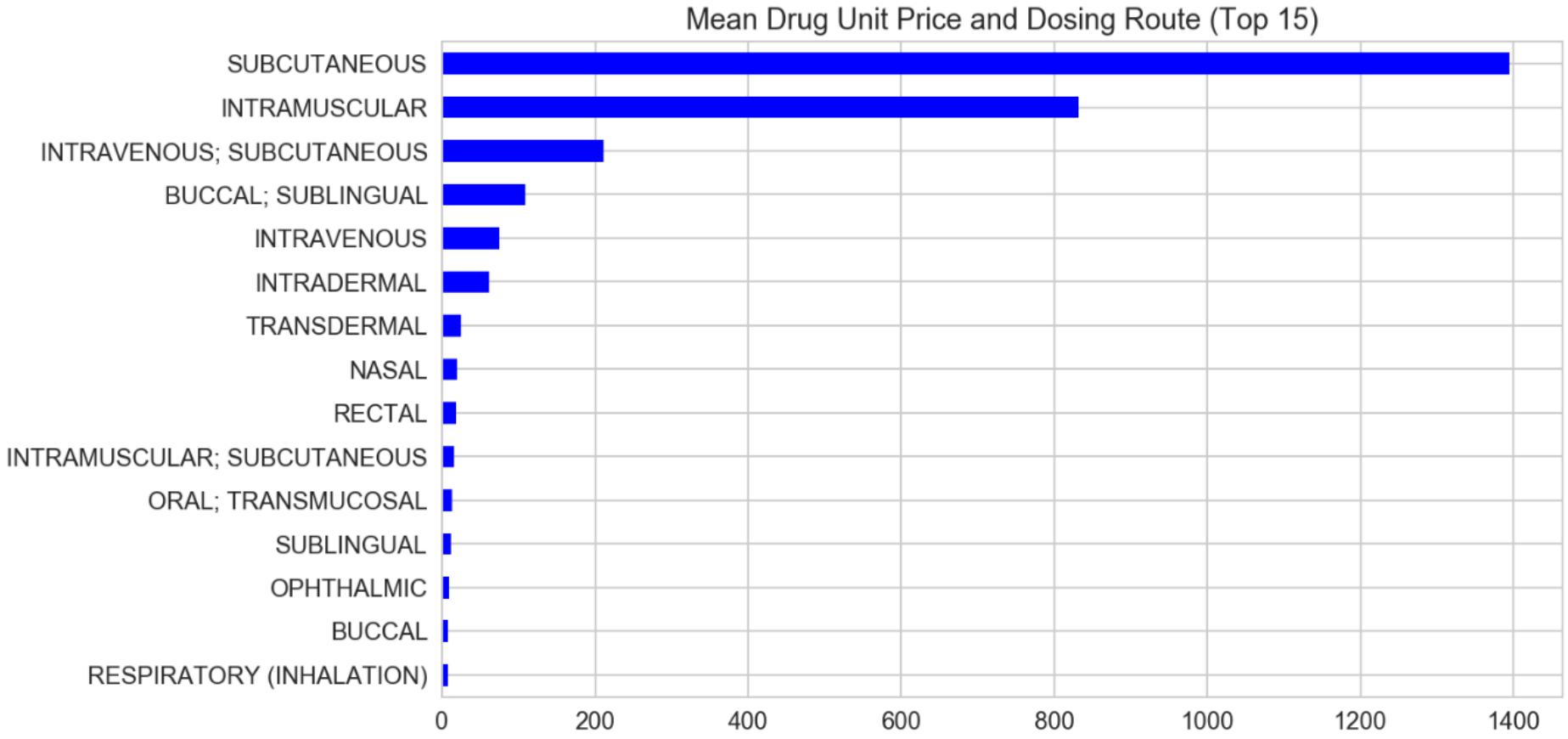
Drug Dosage form has strong impact on drug price

Liquid dosageforms usually are much more expensive than solid and semi-solid dosageform

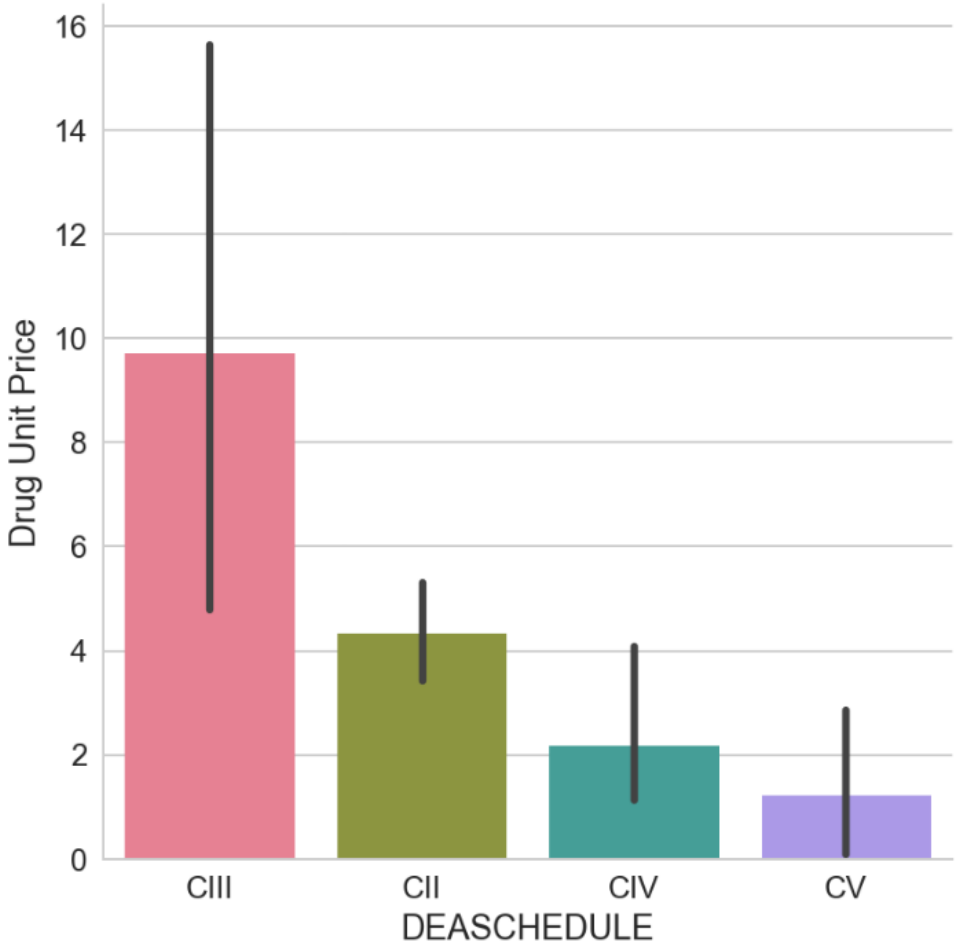


Routes of Administration has strong impact on drug price

Drugs administrated by injection are mostly more expensive than drugs administrated orally



DEA class can make a difference on the price



Schedule I : drugs with no currently accepted medical use and a high potential for abuse: (not covered by Medicaid):

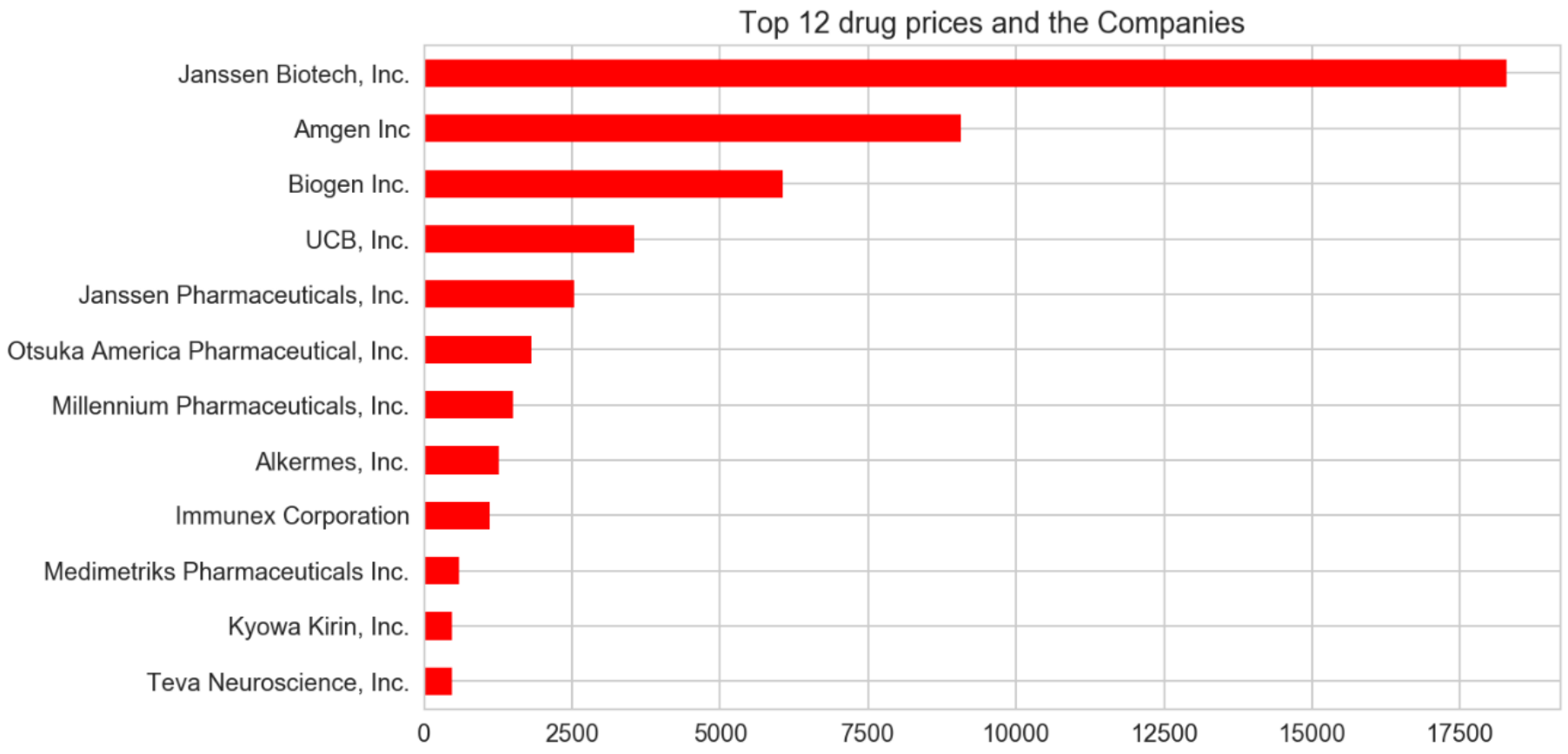
heroin, lysergic acid diethylamide (LSD), marijuana

potential for abuse:

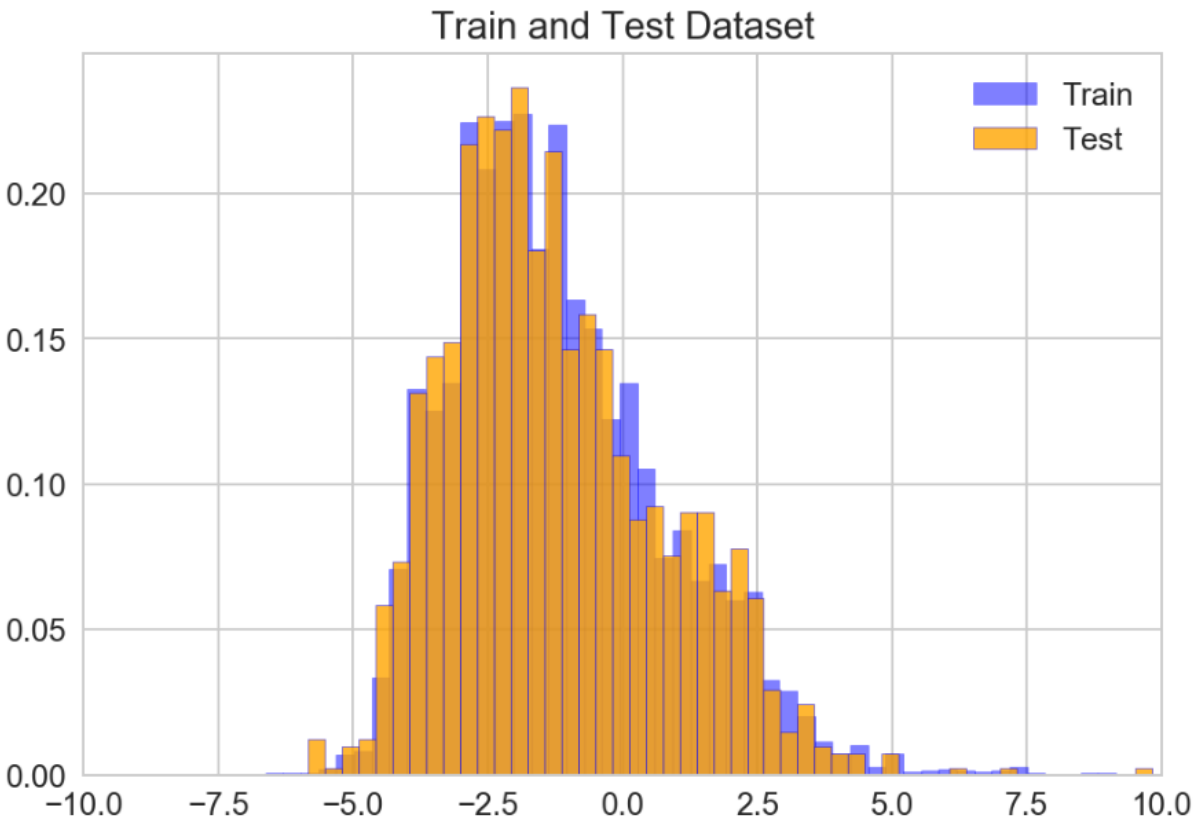
CV<CIV<CIII<CII

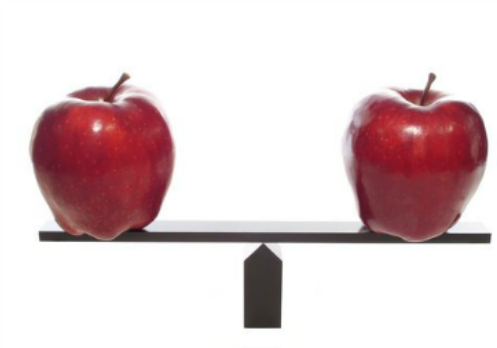
Drug company and price

The top expensive drugs come from top innovative biopharmaceutical companies



Data analysis and modeling





- Unit drug price prediction
- Classification of high (>10 $) and low (< 10$) unit drug price

Regression, Decision Tree, Random Forest, XGBoost

Price Prediction

Price Classification

Predictions of Drug Price

Linear Regression

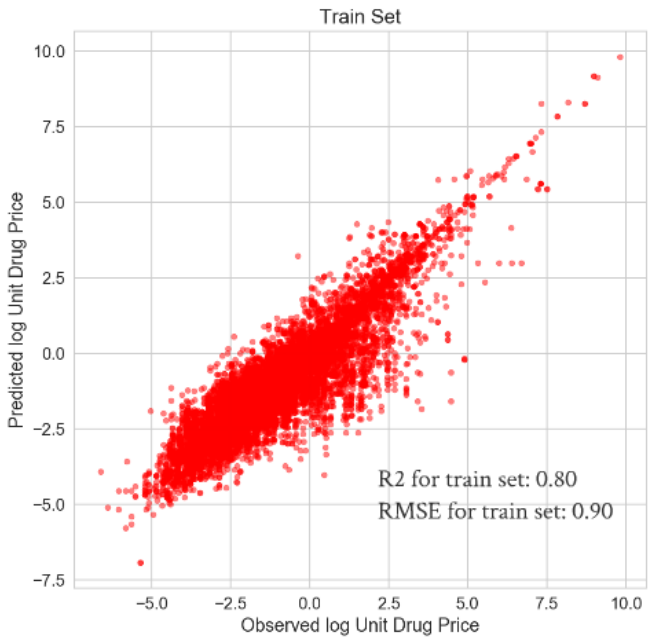
Decision Tree

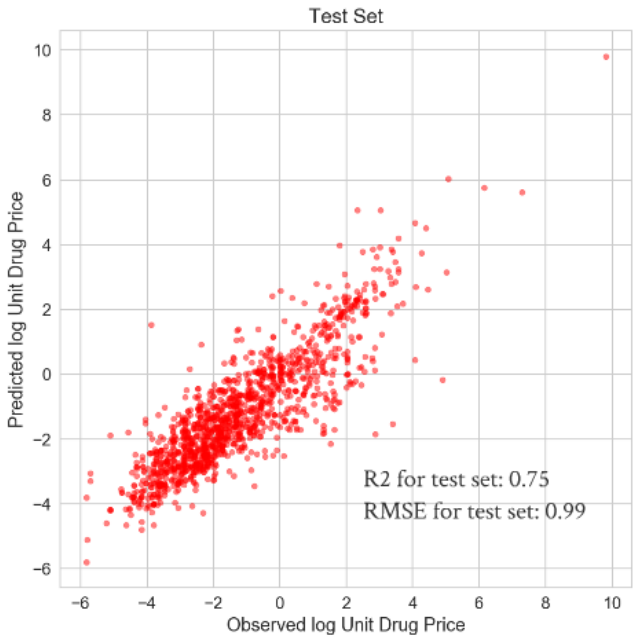
Random Forest

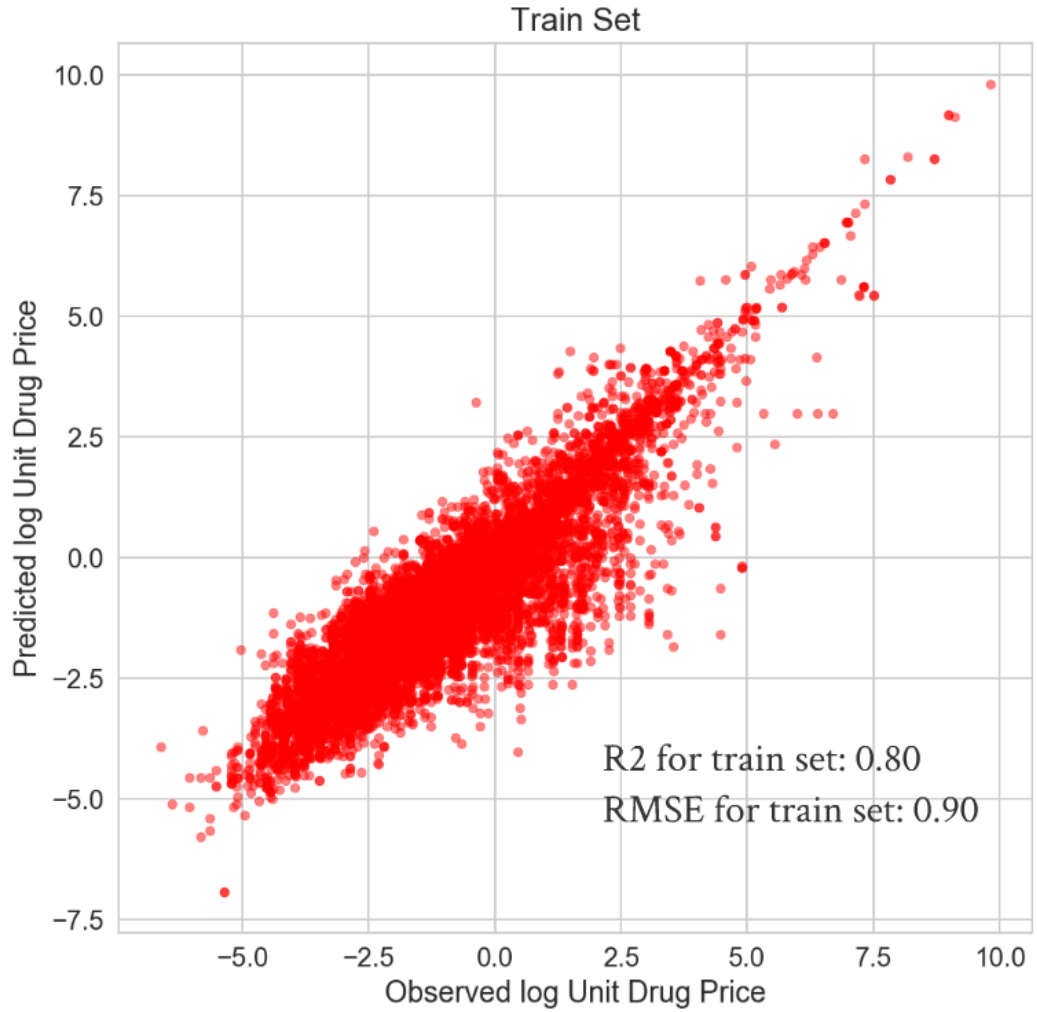
XGboost

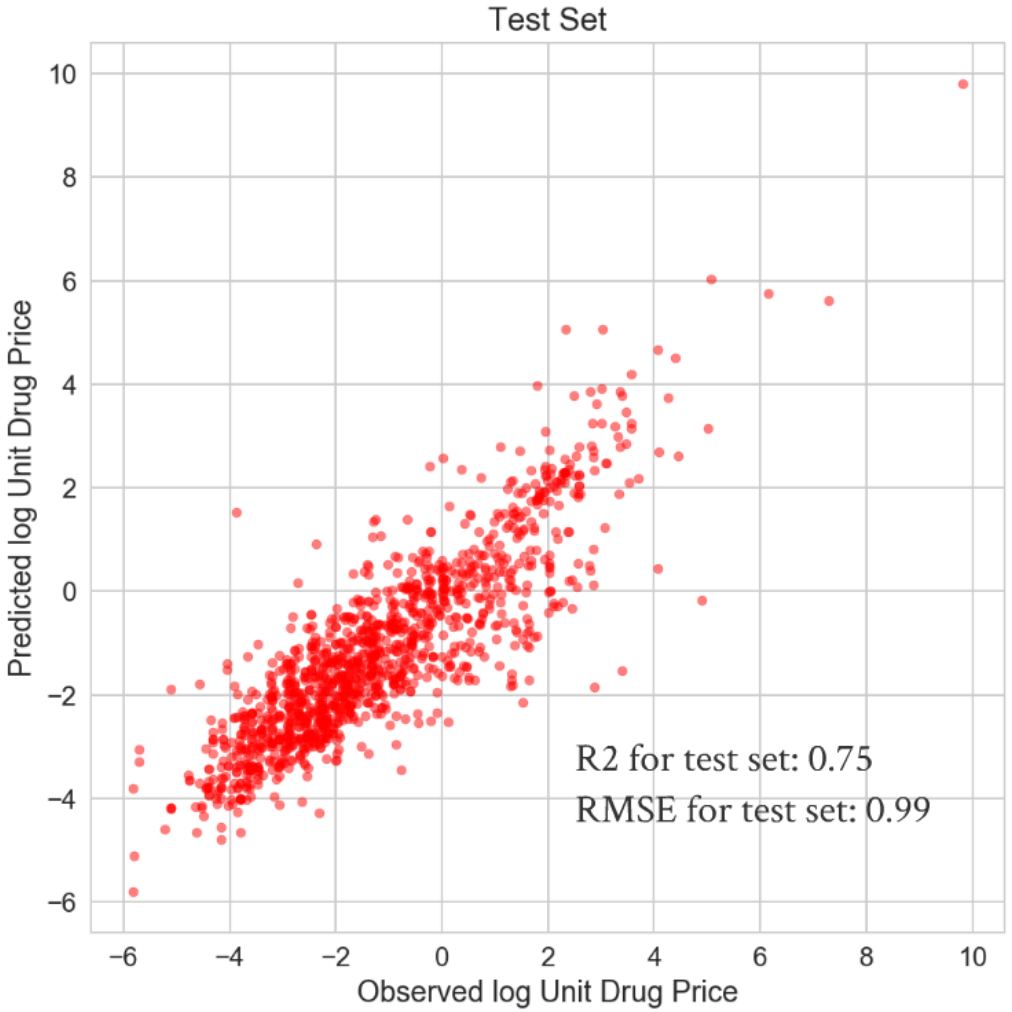
Multiple Linear Regression

The 7 features as Predictors



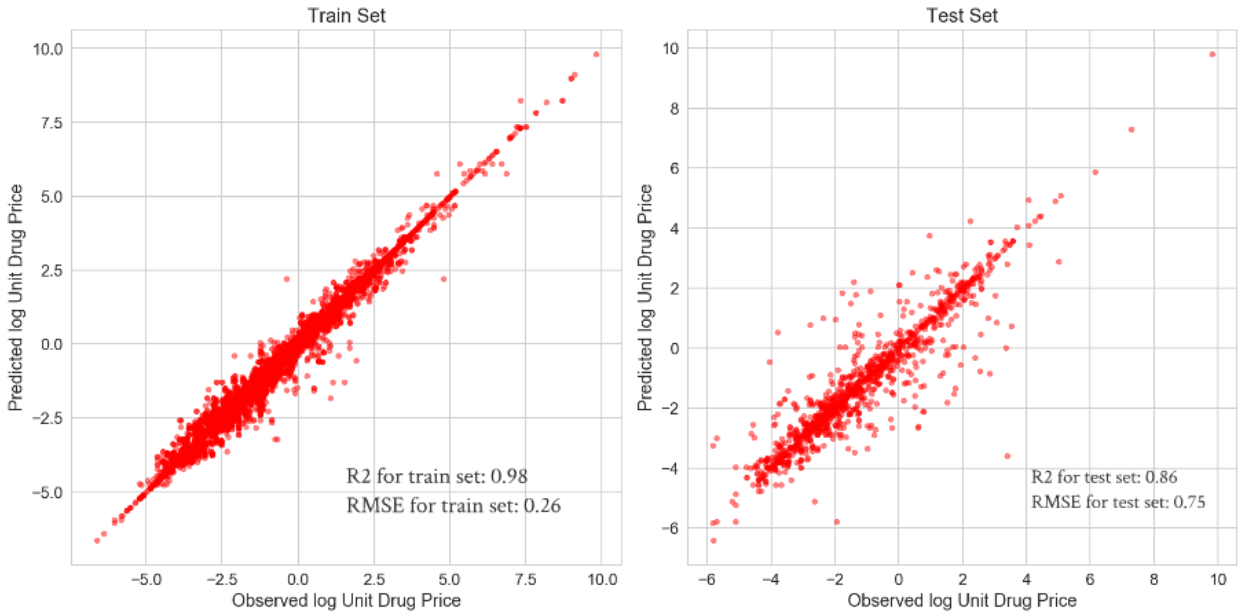


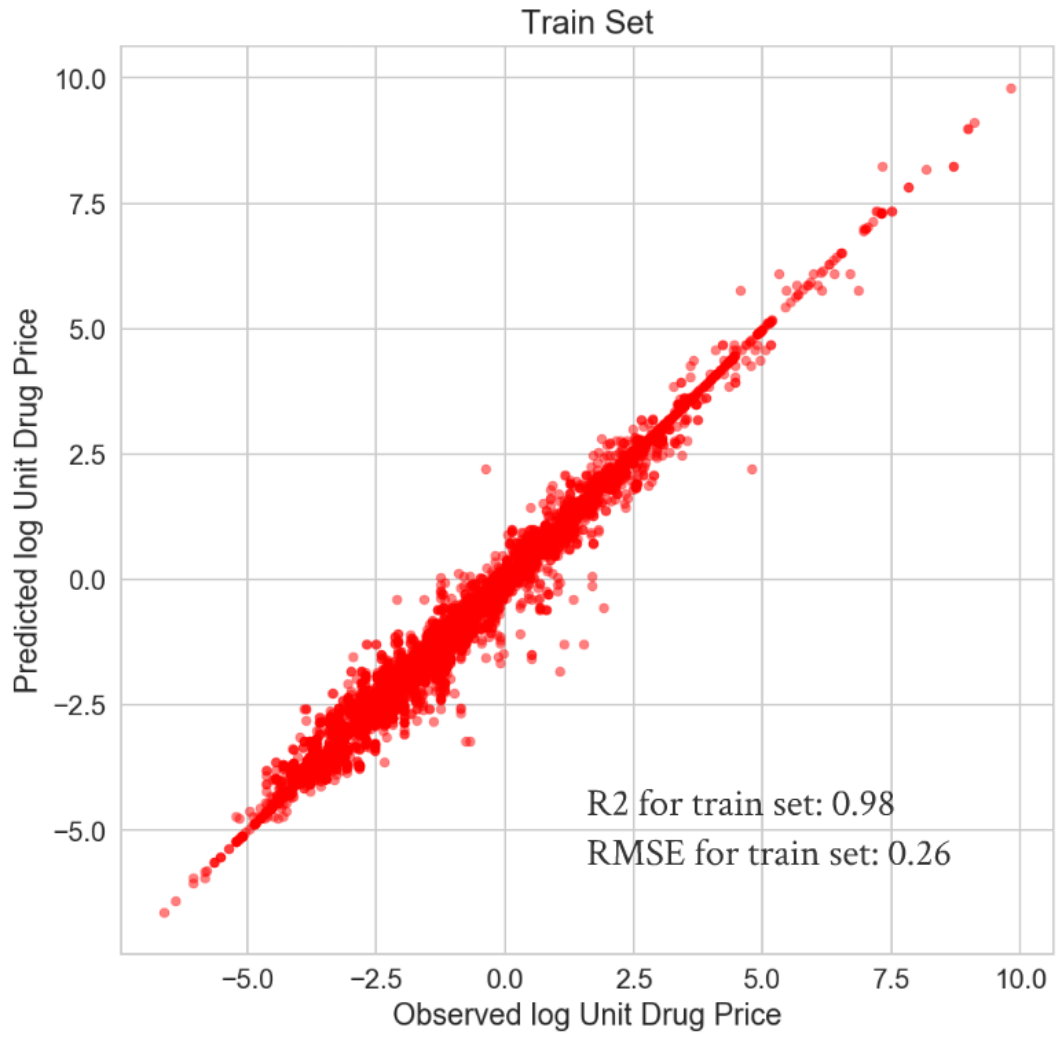

Train Set
Predicted log Unit Drug Price
Observed log Unit Drug Price
R2 for train set: 0.80
RMSE for train set: 0.90

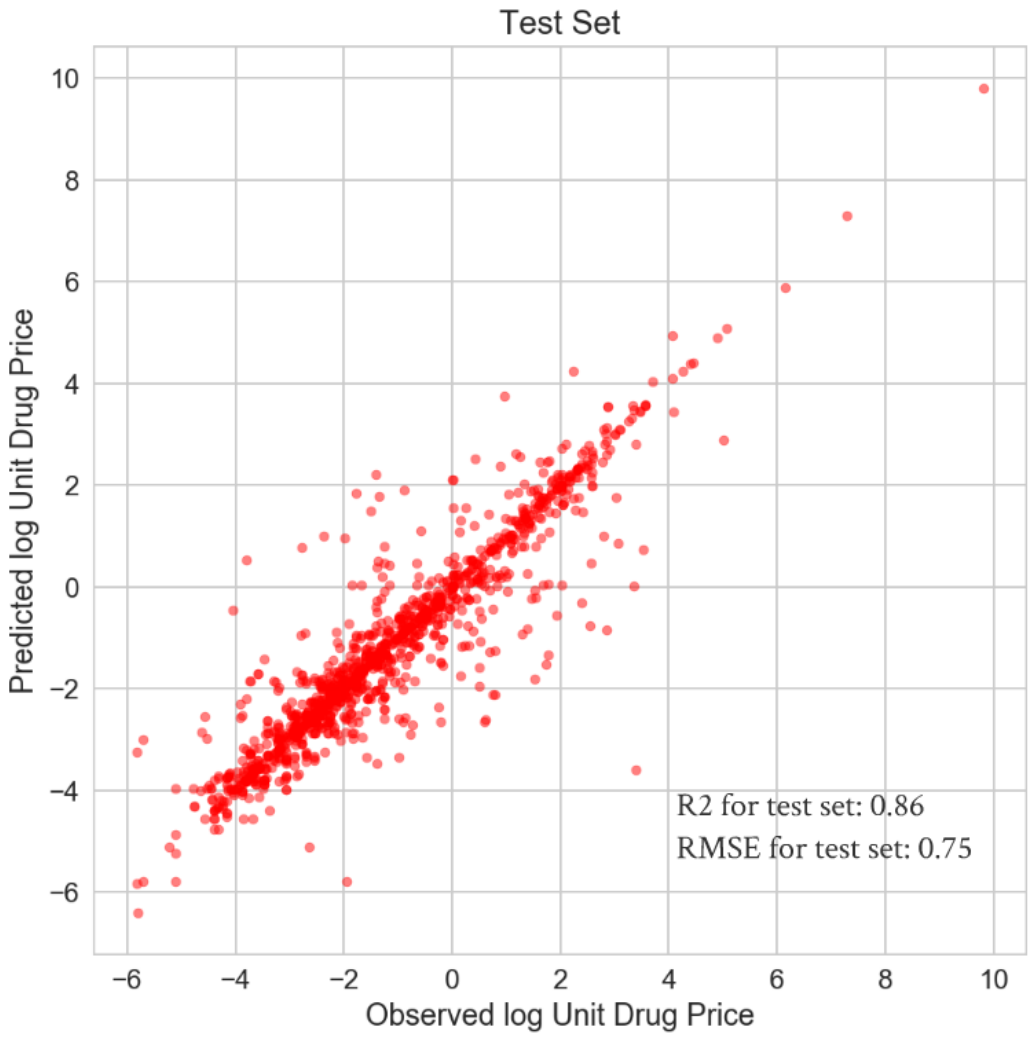

Test Set
Predicted log Unit Drug Price
Observed log Unit Drug Price
R2 for test set: 0.75
RMSE for test set: 0.99

Decision Tree

- Most of the predictors are categorical
- Splits occur at natural locations

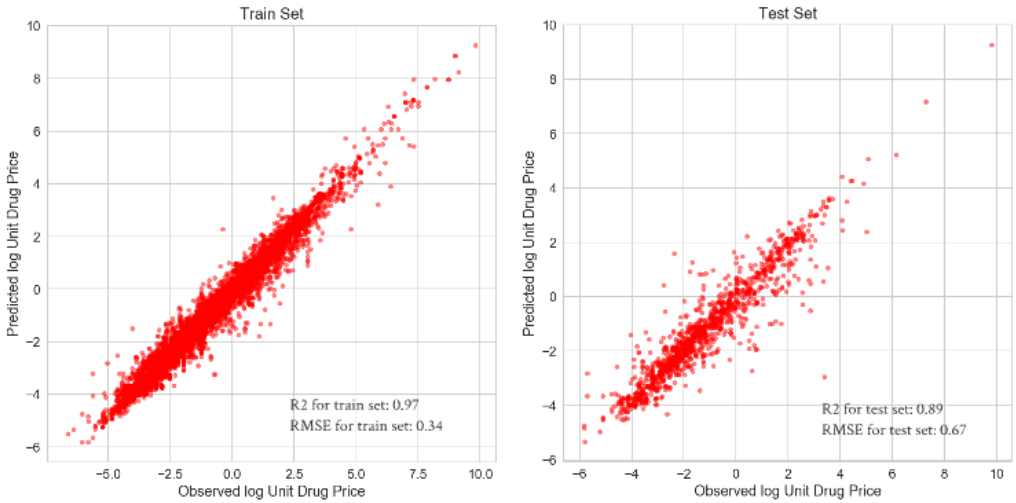


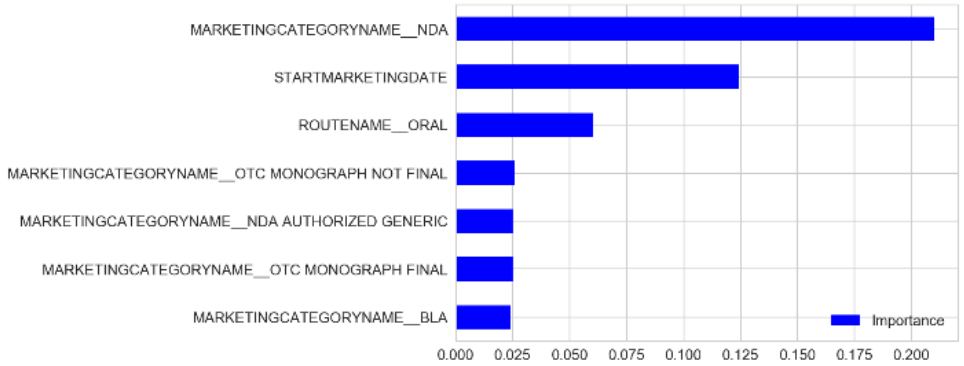

Train Set
Predicted log Unit Drug Price
Observed log Unit Drug Price
R2 for train set: 0.98
RMSE for train set: 0.26


Test Set
Predicted log Unit Drug Price
Observed log Unit Drug Price
R2 for test set: 0.86
RMSE for test set: 0.75

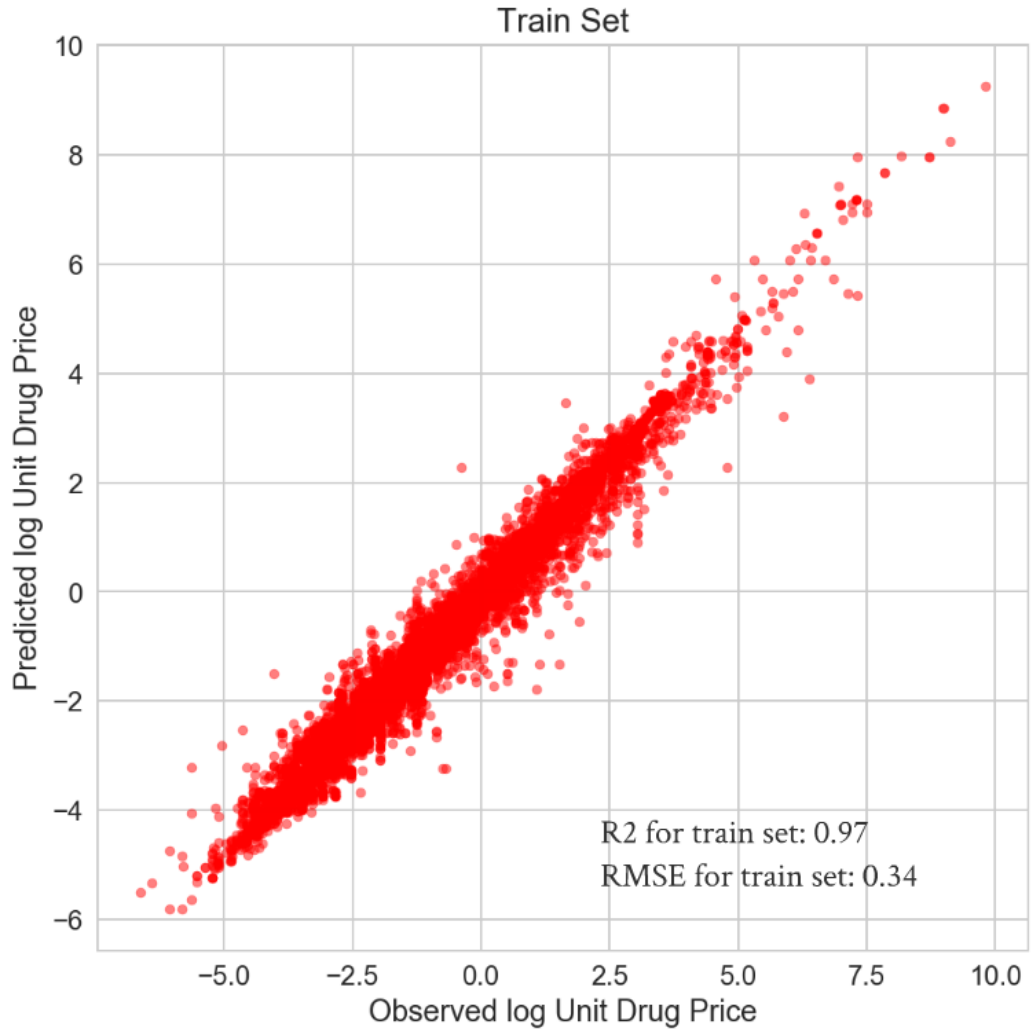
Random Forest

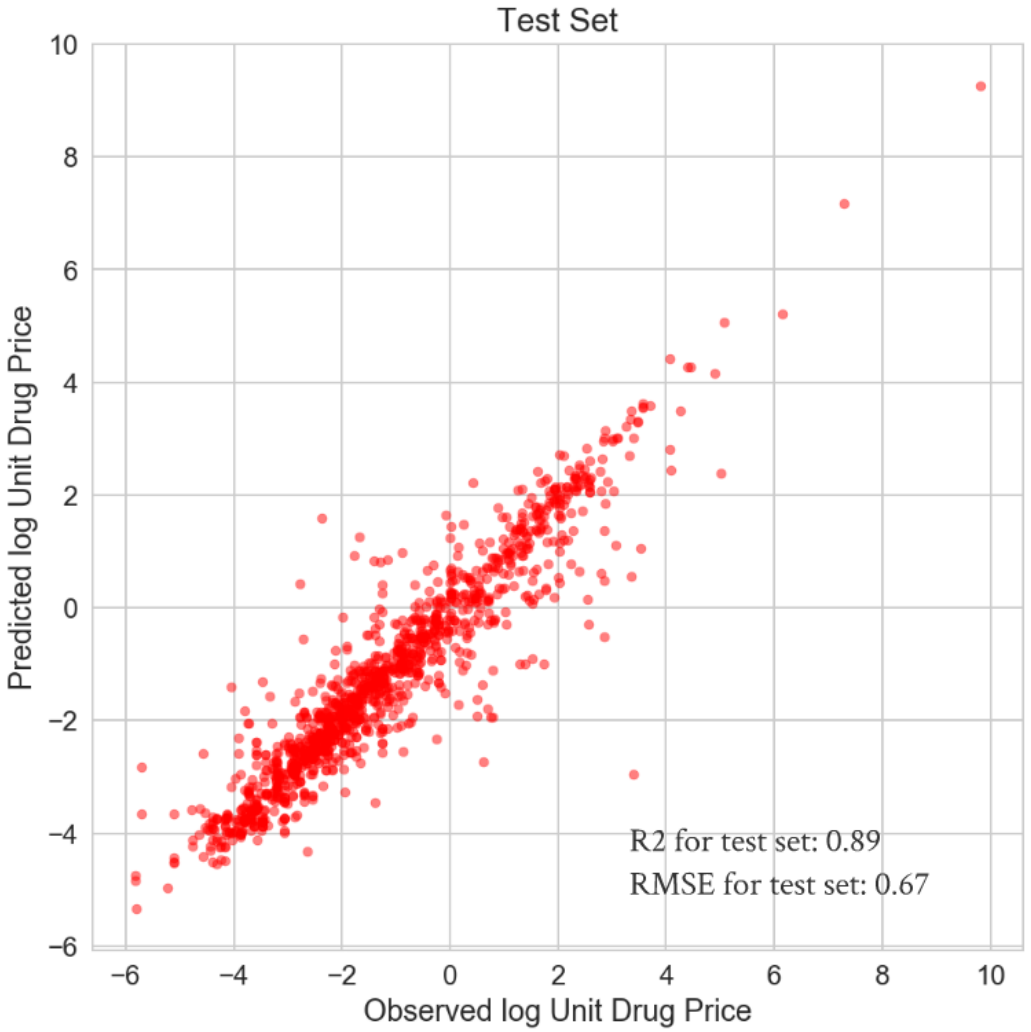
- Prevents overfitting
- Better Prediction

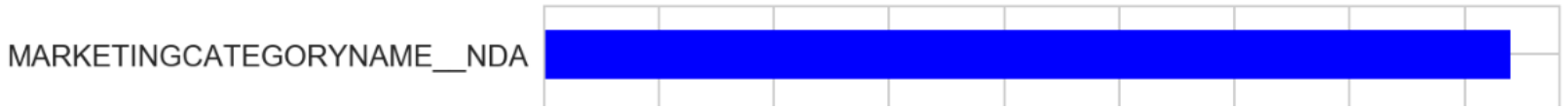


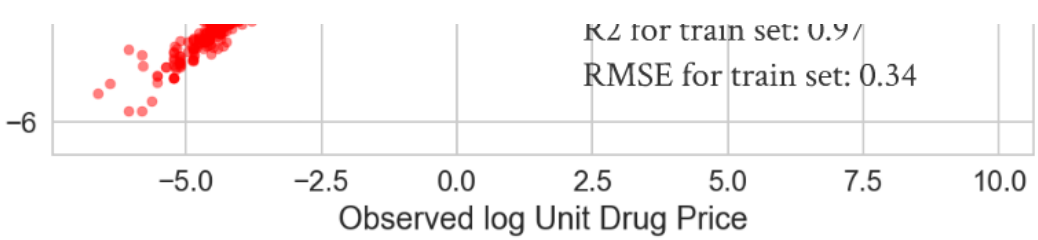


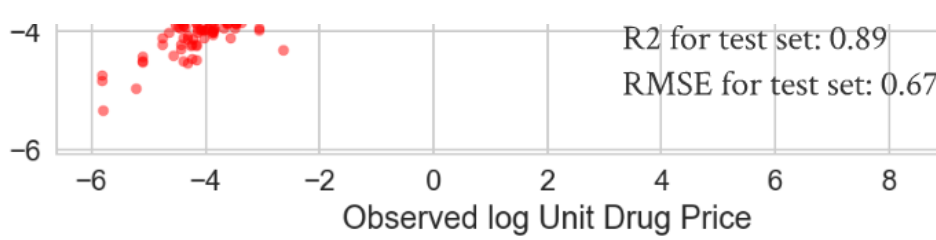
Better Prediction

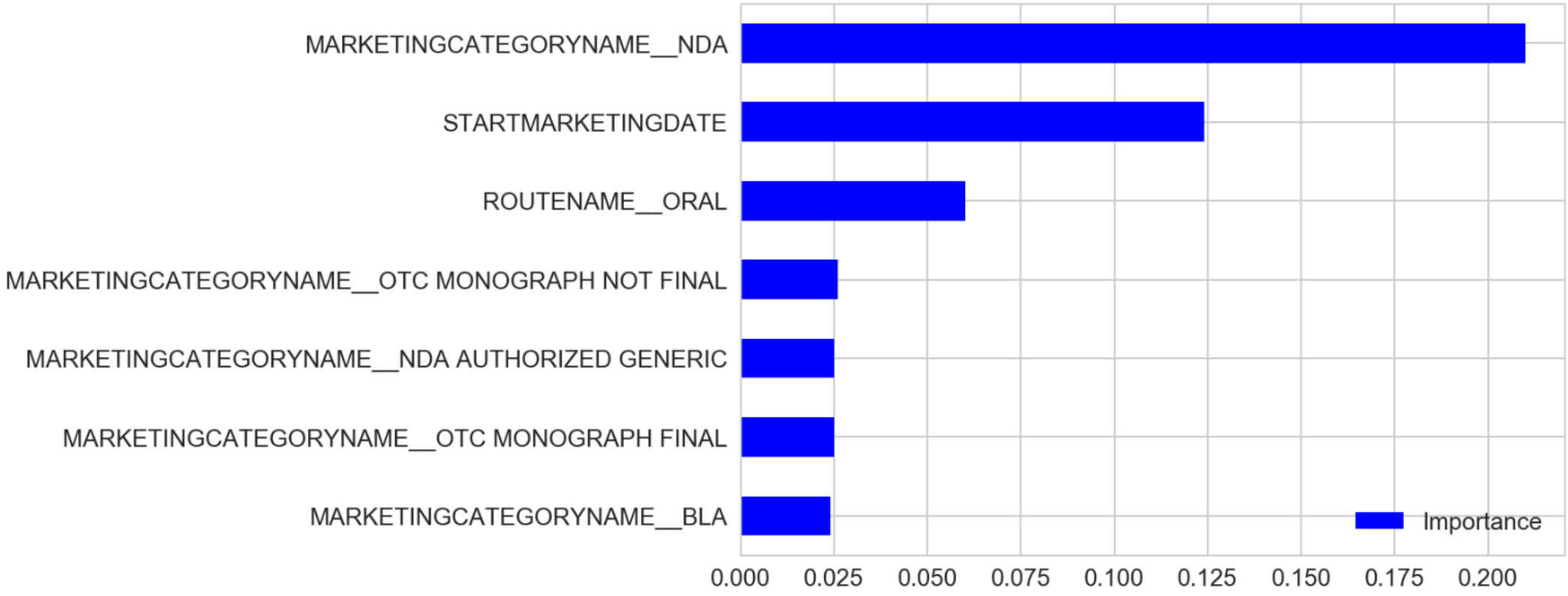






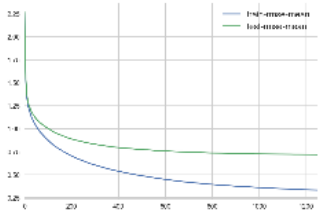

RMSE for train set: 0.34
−6
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
Observed log Unit Drug Price

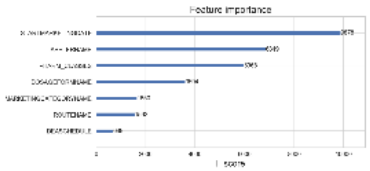

R2 for test set: 0.89
RMSE for test set: 0.67
−4
−6
−6
−4
−2
0
2
4
6
8
Observed log Unit Drug Price

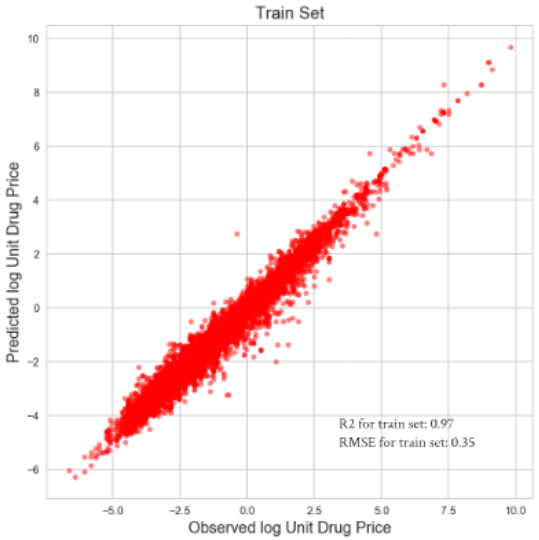

MARKETINGCATEGORYNAME__NDA
STARTMARKETINGDATE
ROUTENAME__ORAL
MARKETINGCATEGORYNAME__OTC MONOGRAPH NOT FINAL
MARKETINGCATEGORYNAME__NDA AUTHORIZED GENERIC
MARKETINGCATEGORYNAME__OTC MONOGRAPH FINAL
MARKETINGCATEGORYNAME__BLA
Importance
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200

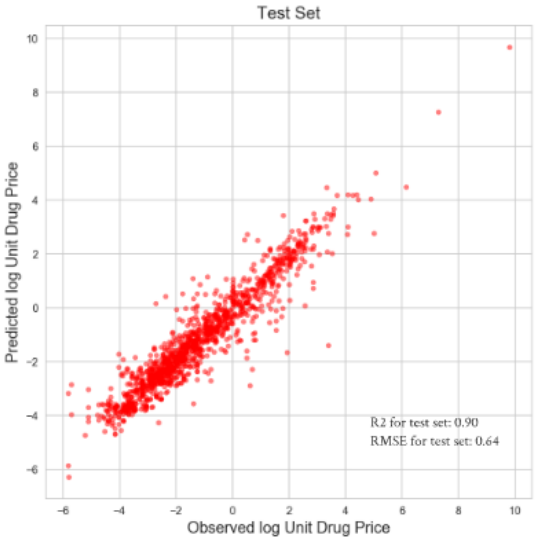
XGBoost

- Performs slightly better than Random Forest
- Cross Validation, 5-fold

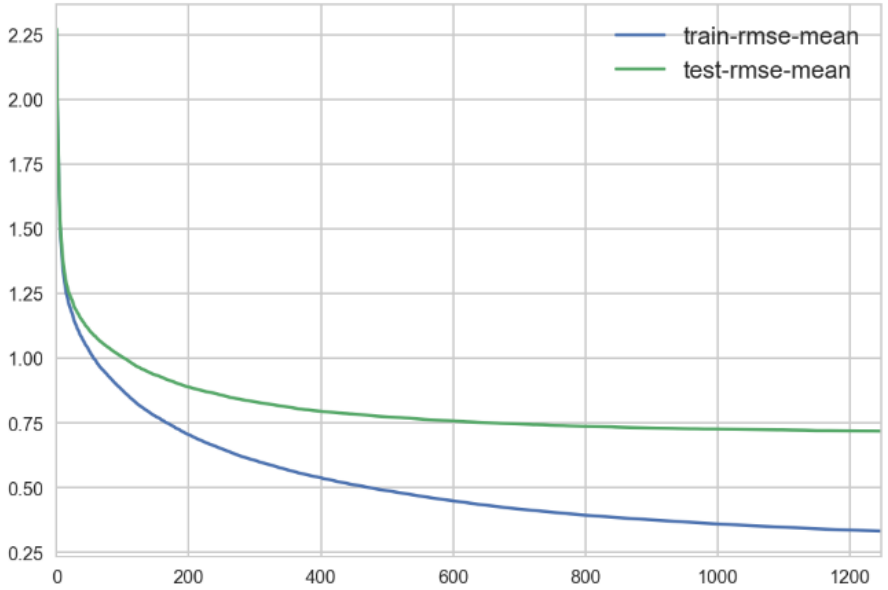


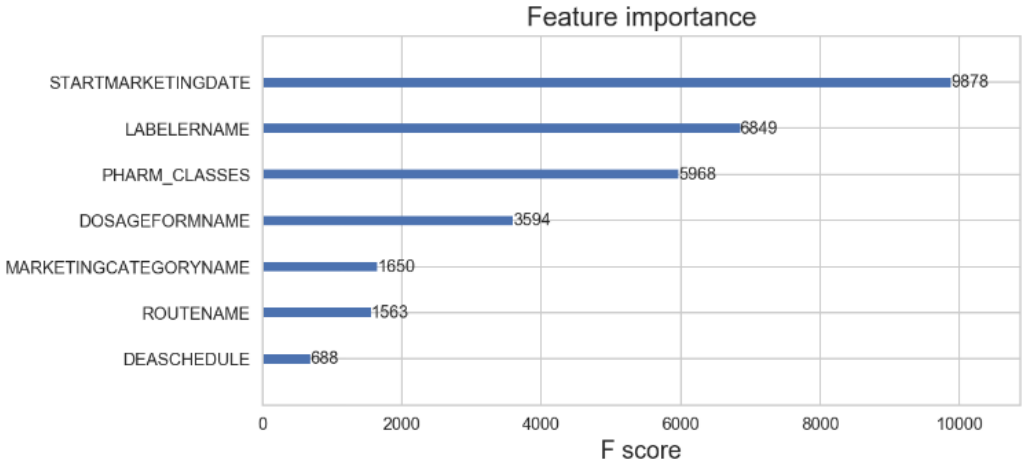






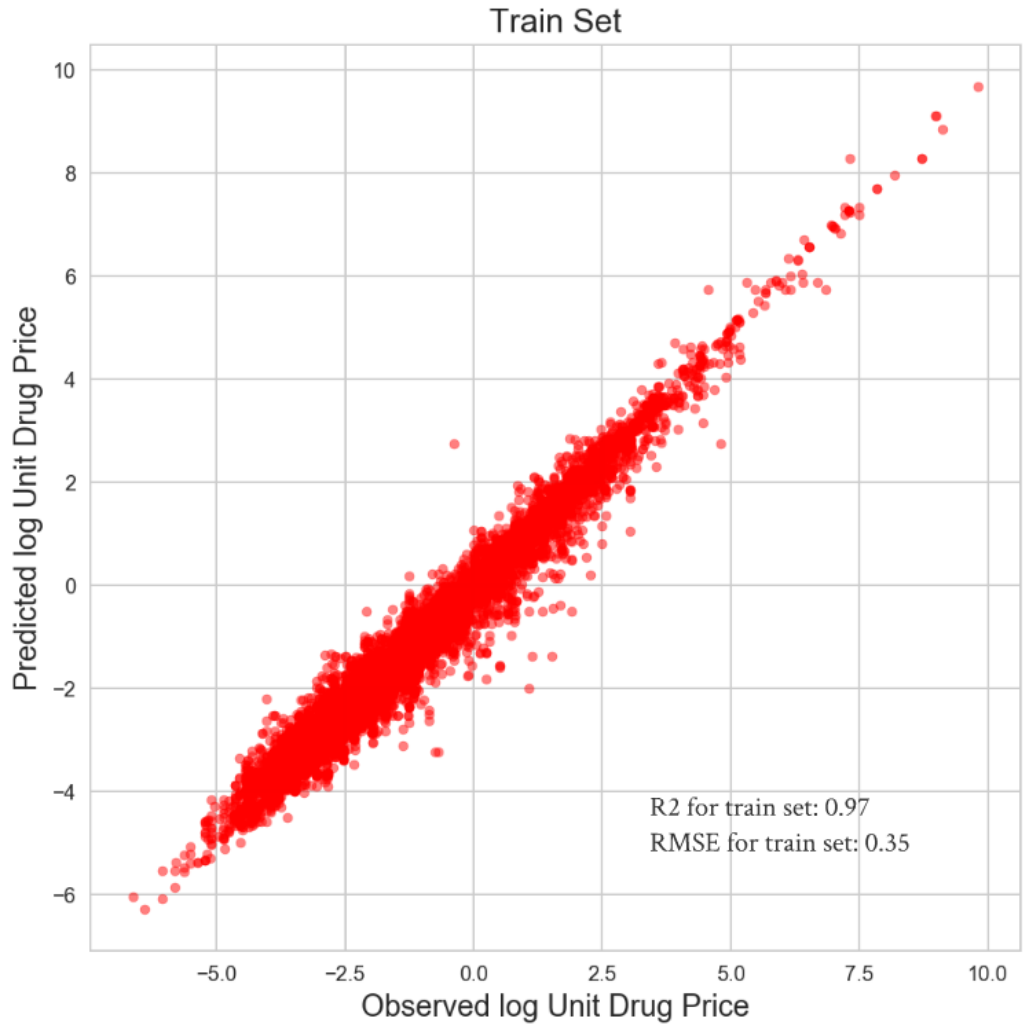
- Cross Validation, 5-fold

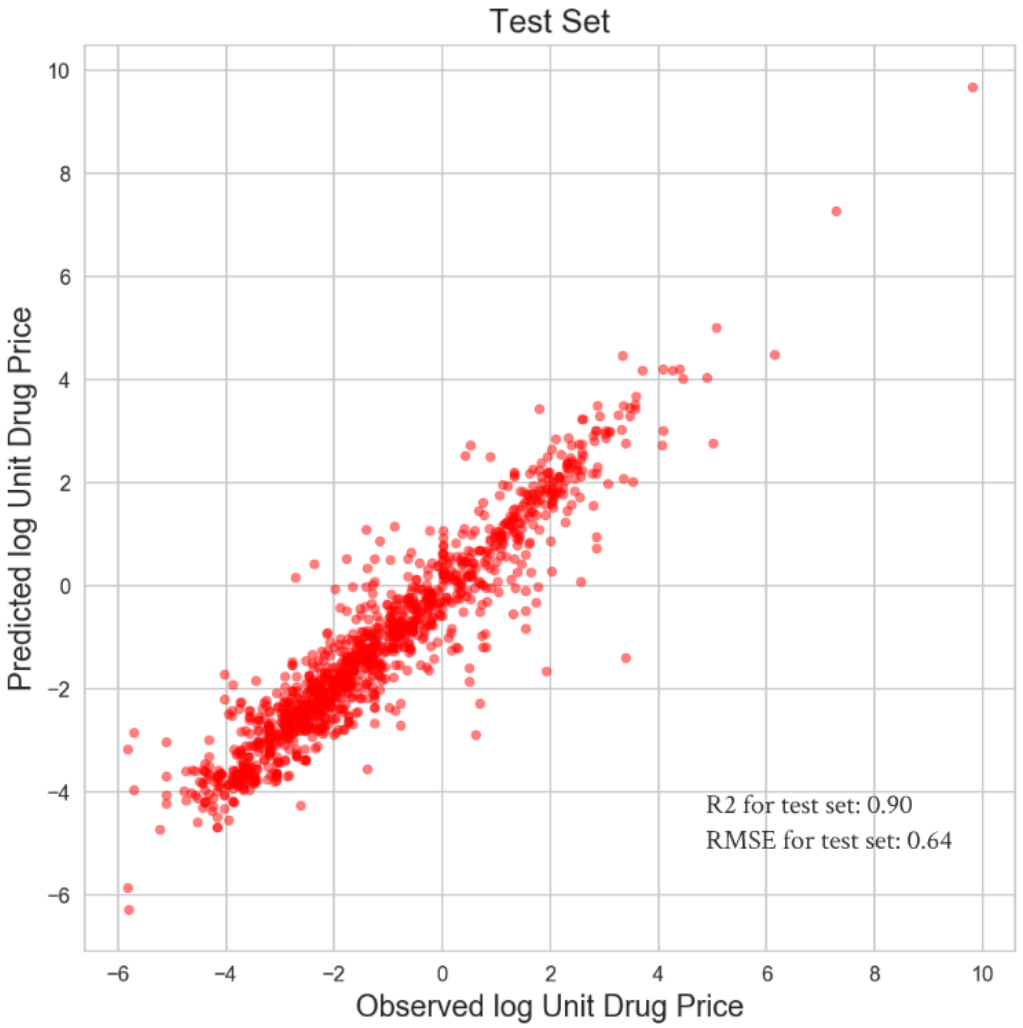




Train Set

Test Set


Train Set
Predicted log Unit Drug Price
Observed log Unit Drug Price
R2 for train set: 0.97
RMSE for train set: 0.35


Test Set
Predicted log Unit Drug Price
Observed log Unit Drug Price
R2 for test set: 0.90
RMSE for test set: 0.64

Drug Price Classification

Logistic Regression

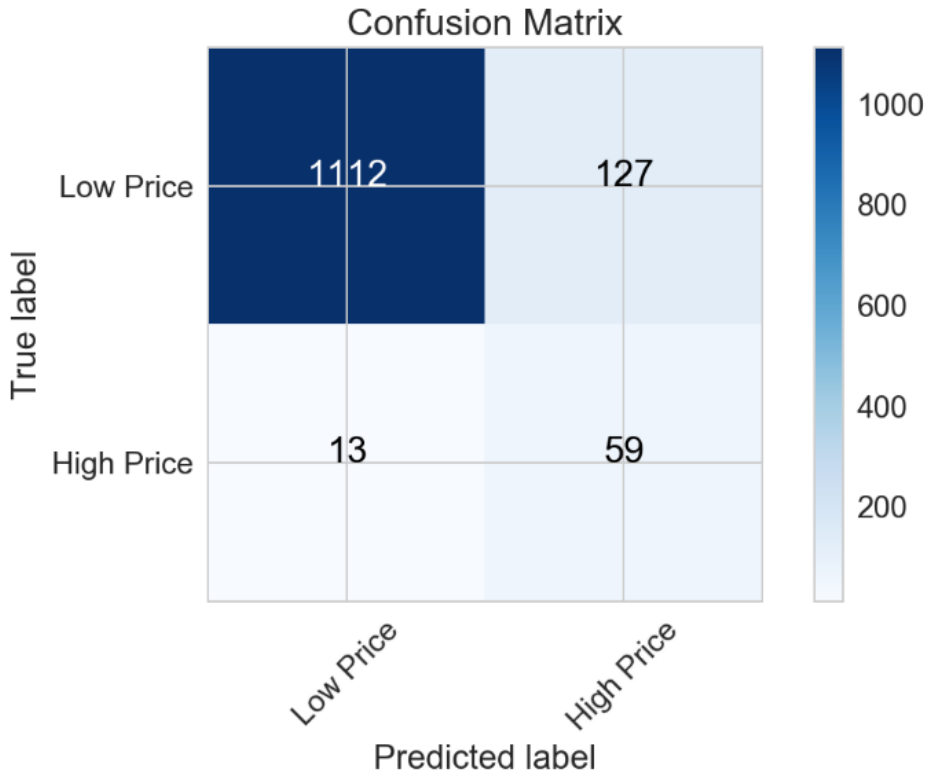
Decision Tree

Random Forest

XGboost

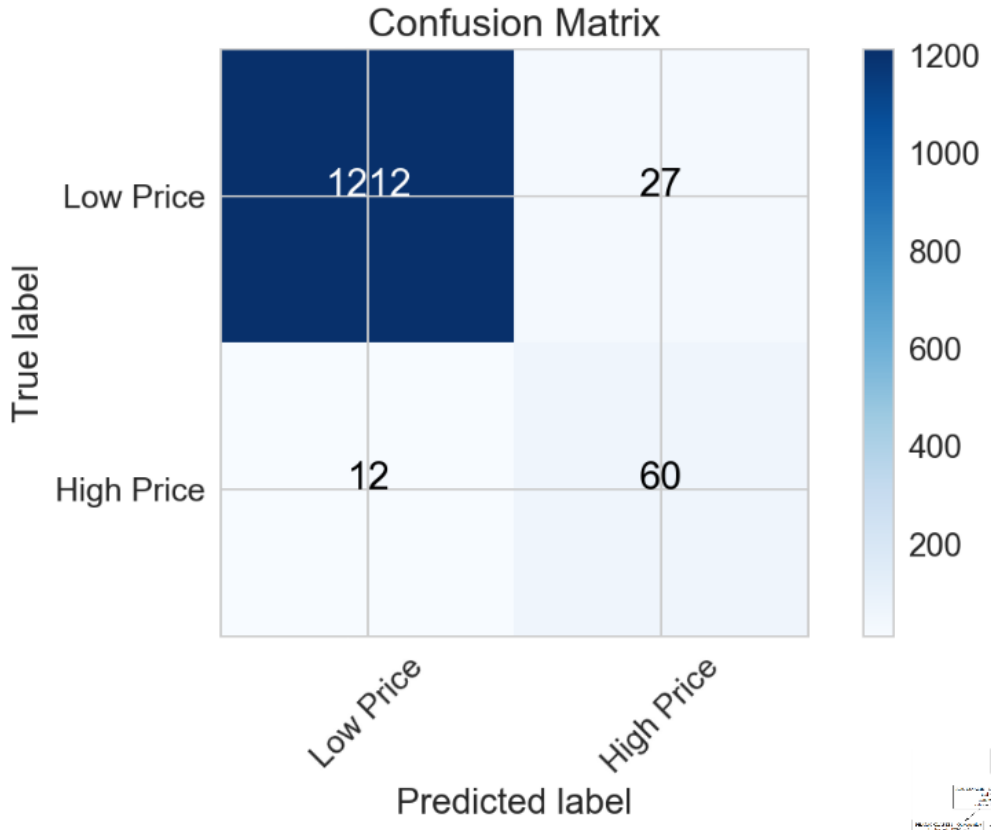
Logistic Regression

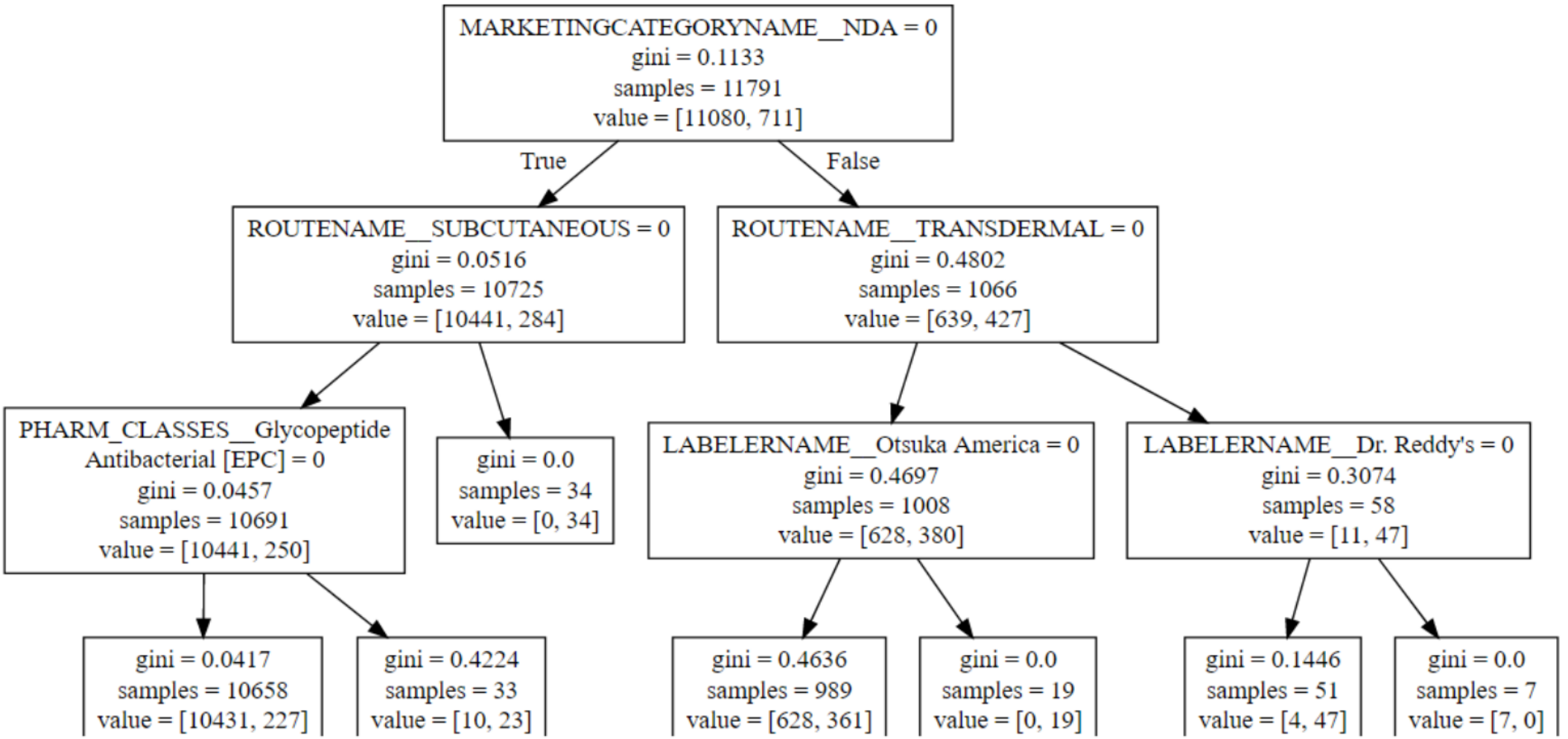
- Poor precision on high price drugs
- Accuracy Score: 0.92
- Low Price Precision: 0.99
- High Price Precision: 0.32
- Stratified sampling of cross validation, 5-fold



Decision Tree

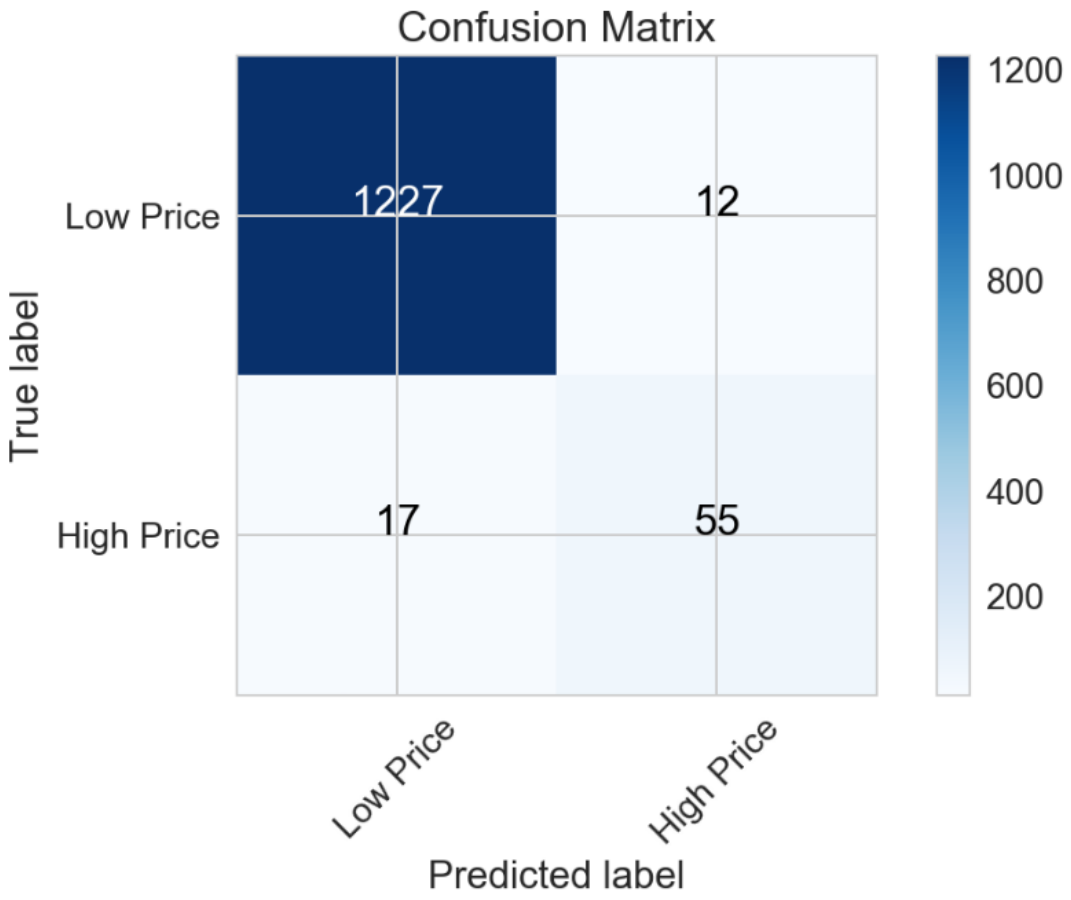
- Improved precision on high price drugs
- Accuracy Score: 0.97
- Low Price Precision: 0.99
- High Price Precision: 0.69
- Stratified sampling of cross validation




MARKETINGCATEGORYNAME__NDA = 0
gini = 0.1133
samples = 11791
value = [11080, 711]
True
False
ROUTENAME__SUBCUTANEOUS = 0
gini = 0.0516
samples = 10725
value = [10441, 284]
ROUTENAME__TRANSDERMAL = 0
gini = 0.4802
samples = 1066
value = [639, 427]
PHARM_CLASSES__Glycopeptide Antibacterial [EPC] = 0
gini = 0.0457
samples = 10691
value = [10441, 250]
gini = 0.0
samples = 34
value = [0, 34]
LABELERNAME__Otsuka America = 0
gini = 0.4697
samples = 1008
value = [628, 380]
LABELERNAME__Dr. Reddy's = 0
gini = 0.3074
samples = 58
value = [11, 47]
gini = 0.0417
samples = 10658
value = [10431, 227]
gini = 0.4224
samples = 33
value = [10, 23]
gini = 0.4636
samples = 989
value = [628, 361]
gini = 0.0
samples = 19
value = [0, 19]
gini = 0.1446
samples = 51
value = [4, 47]
gini = 0.0
samples = 7
value = [7, 0]

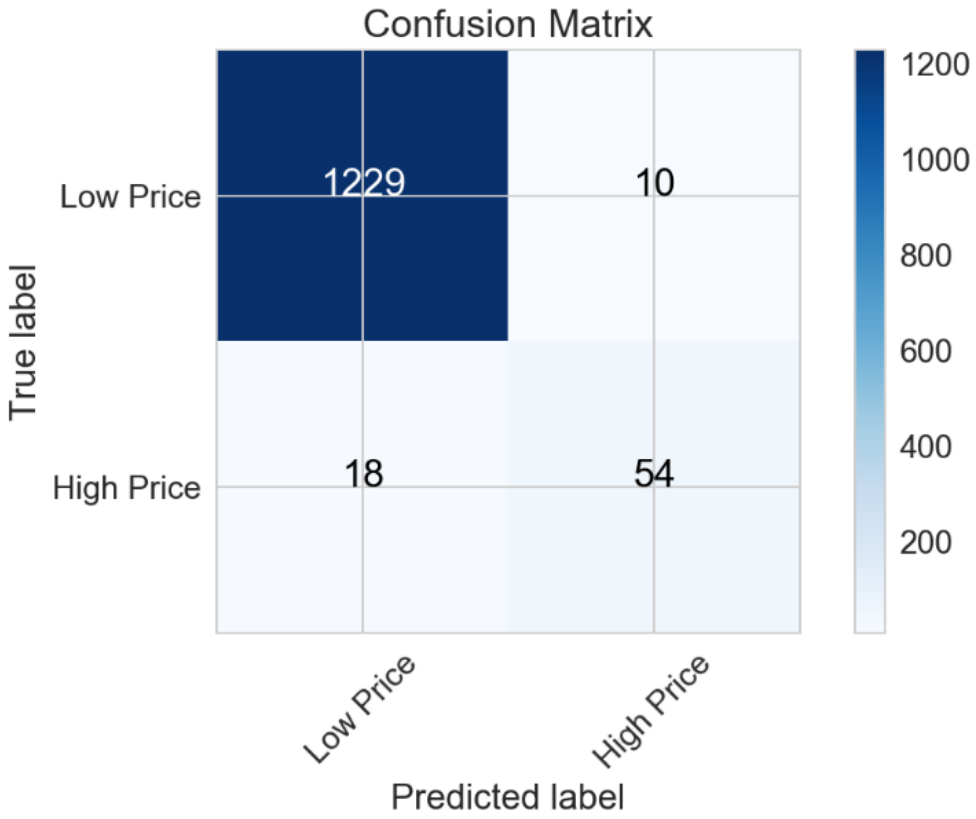
Random Forest

- Much improved estimation on high price drugs
- Accuracy Score: 0.98
- Low Price Precision: 0.99
- High Price Precision: 0.82



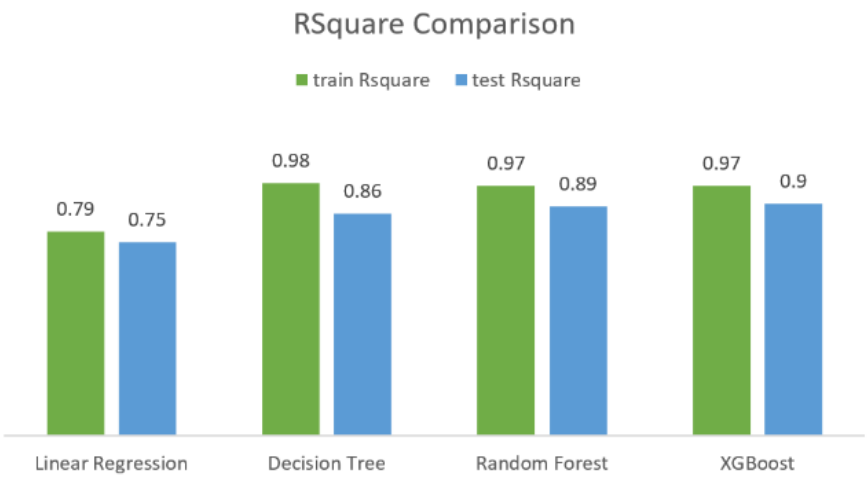
XGBoost

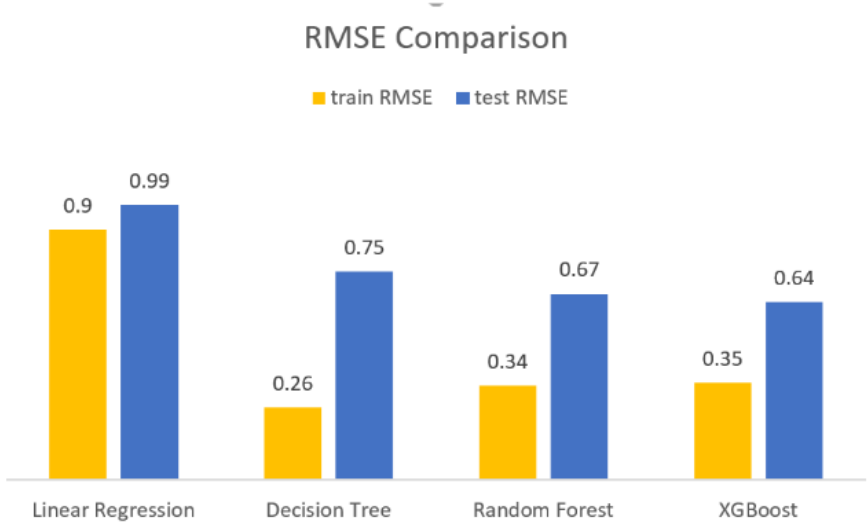
- Best result among all models
- Accuracy Score: 0.98
- Low Price Precision: 0.99
- High Price Precision: 0.84



Comparisons: Machine Learning Models

Price Prediction Models





Price Classfication Models

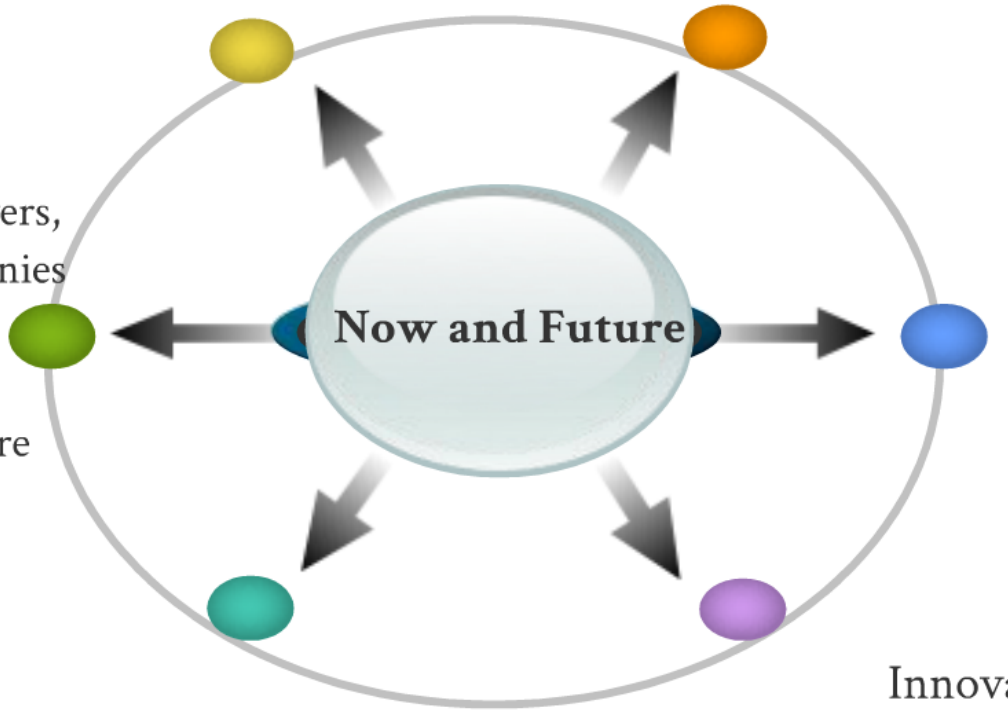
Conclusions

Prediction and classification models
Reasonably robust

Regression<Tree<Random Forest<=XGBoost

Significant values:

- patients, prescribers, payers, medical insurance companies
- new drug development evaluation
- target selection for venture capital investment.



Expand the dataset
Optimize the model
supply-demand, economics, policy, regulatory variables, time series analysis

Innovative work
for publication

Apply to other fields and products

Acknowledgment

The NYCDSA Team

The Fellow Bootcamp Students

The Core Team: Stay Healthy, Live Longer!

Questions