



PREDICTING SUBSTANCE USE TREATMENT COMPLETION WITH DATA SCIENCE

KEVIN PENG



Table of Contents

- Problem Statement
- Outside Information
 - Deaths of Despair
- Dataset
- EDA
- Feature Selection
- Feature Engineering
- Model Performance
 - Basic
 - Hypertuned
- Findings
- Recommendations
- Next Steps
- Resources
- Public Service Announcement – XGBoost



Problem Statement

Kevin Peng was hired by the ASAC (Association of Substance Abuse Clinicians) IPA to construct a model to help their members make predictions on whether or not a patient will complete treatment. The goal is to use the model to help with data-driven decision-making and resource allocation.

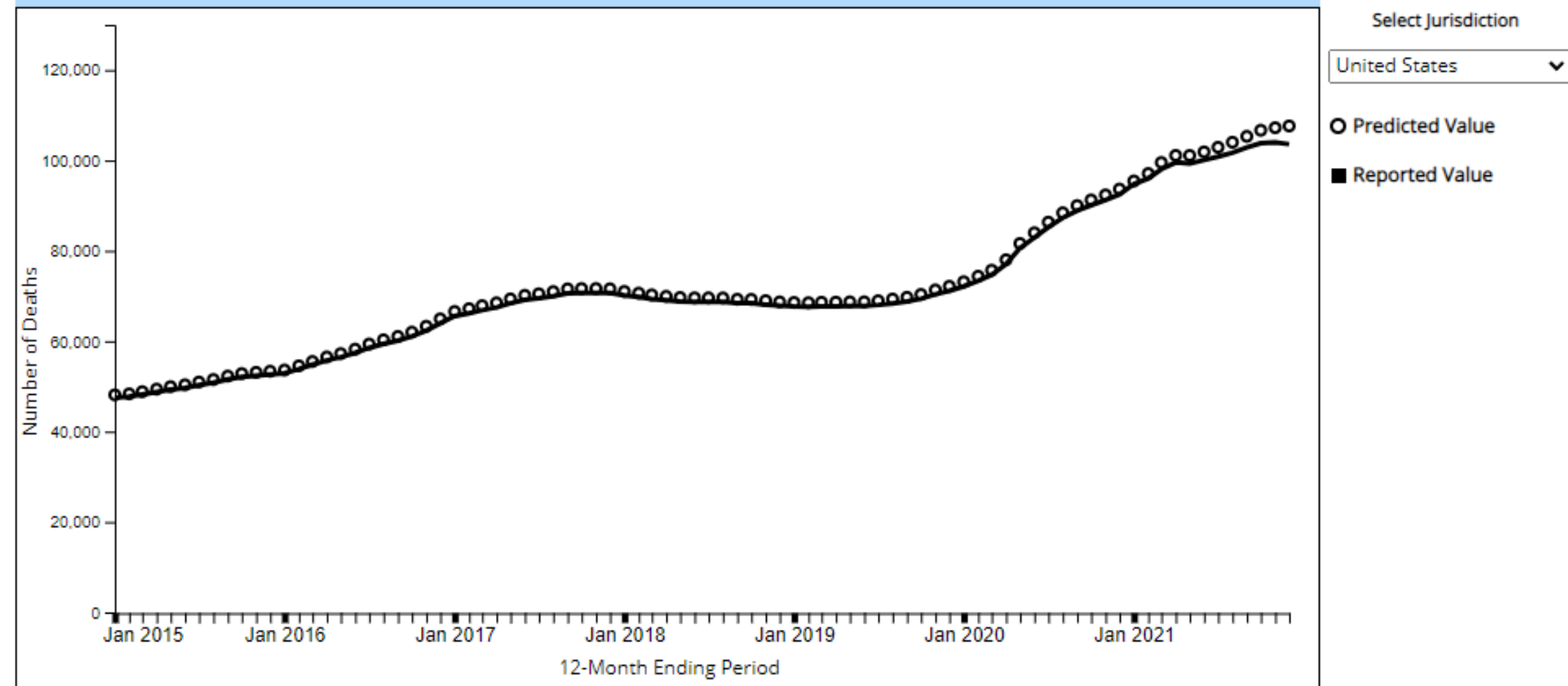
Outside Information

Source: CDC

12 Month-ending Provisional Number and Percent Change of Drug Overdose Deaths

Based on data available for analysis on: May 01, 2022

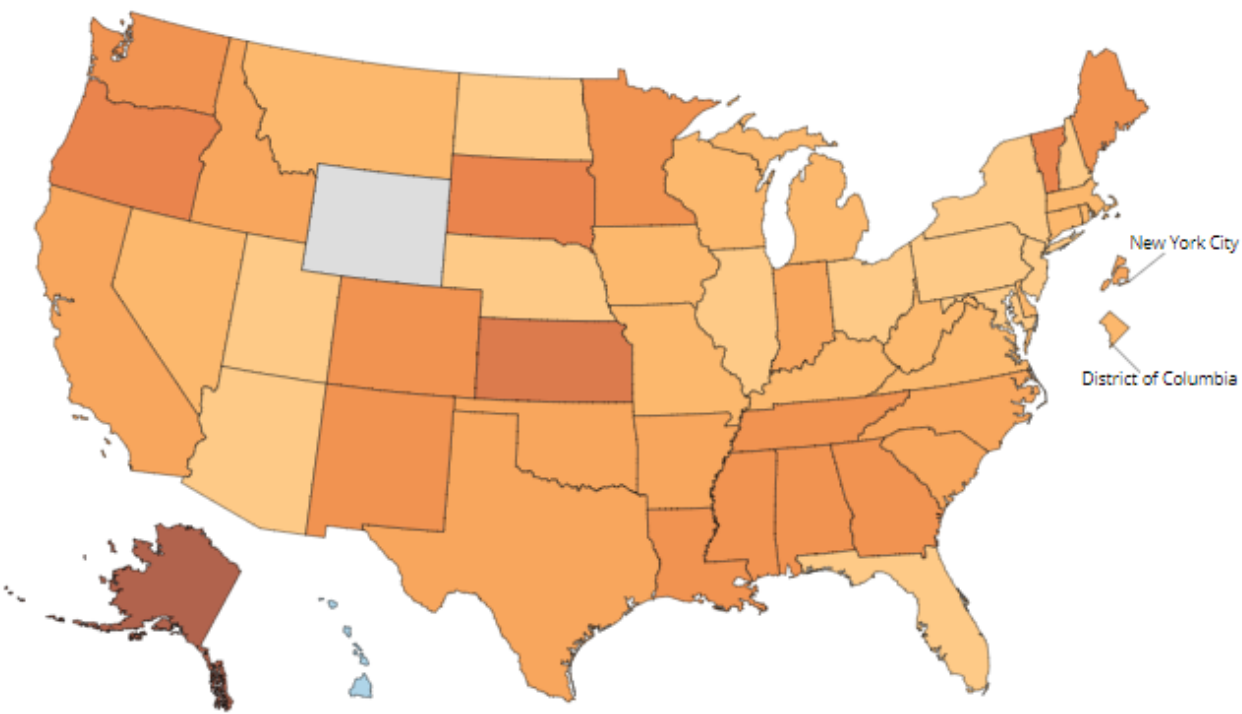
Figure 1a. 12 Month-ending Provisional Counts of Drug Overdose Deaths: United States



Outside Information

Source: CDC

Figure 1b. Percent Change in Predicted 12 Month-ending Count of Drug Overdose Deaths, by Jurisdiction: December 2020 to December 2021



Legend for Percent Change in Drug Overdose Deaths Between 12-Month Ending Periods

-1.8 75.3

Select predicted or reported number of deaths

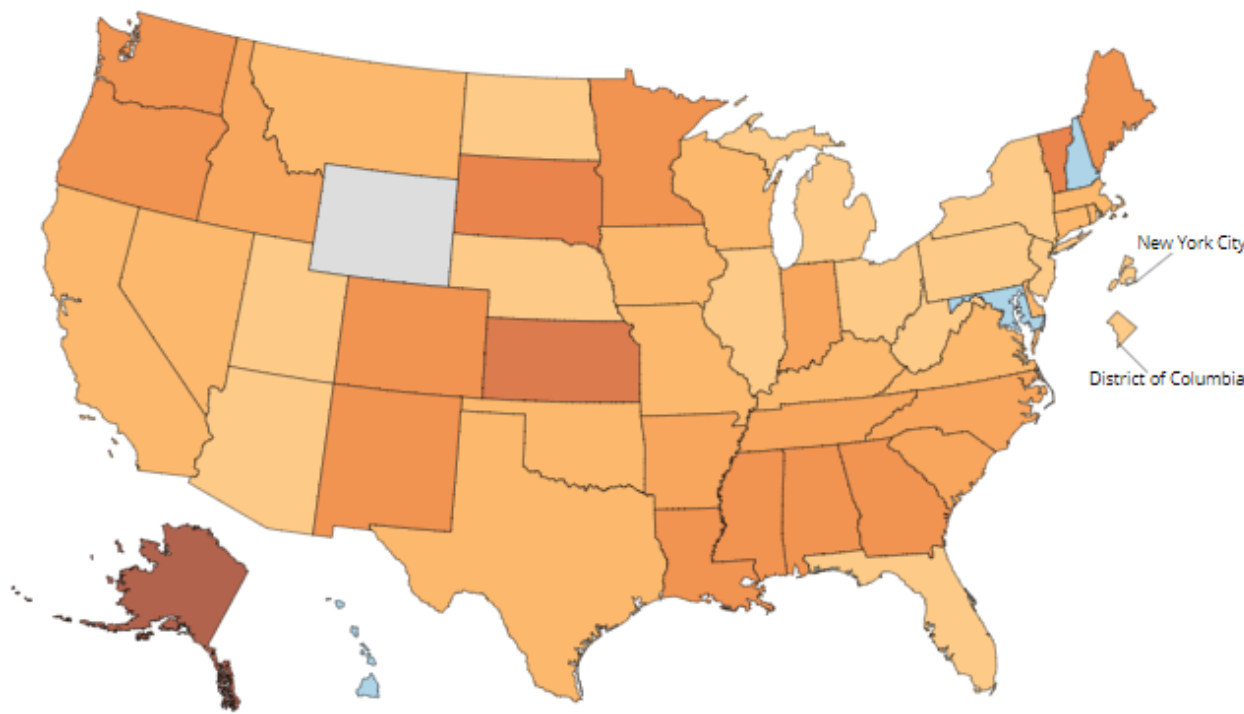
☒ Predicted

☐ Reported

Percent Change for United States

14.9 ▲

Figure 1b. Percent Change in Reported 12 Month-ending Count of Drug Overdose Deaths, by Jurisdiction: December 2020 to December 2021



Legend for Percent Change in Drug Overdose Deaths Between 12-Month Ending Periods

-9.1 73.3

Select predicted or reported number of deaths

☐ Predicted

☒ Reported

Percent Change for United States

12.0 ▲

Deaths of Despair

01 Suicide

02 Alcohol

03 Overdose

Dataset

- Original Data
 - 1722503 rows × 76 columns
 - Categorical/Ordinal
- Data Cleanup
 - Converted Nulls
 - Drop all NA - 293 Rows Remain (1.7%)
 - Removed Rows - 1590685 Remain (92%)
 - Removed Columns - 17 Remain (22%)
- Storage
 - SQLite database



Dataset

- Detoxification, 24-hour service, hospital inpatient:

24 hours per day medical acute care services in hospital setting for detoxification of persons with severe medical complications associated with withdrawal.

- Detoxification, 24-hour service, free-standing residential:

24 hours per day services in non-hospital setting providing for safe withdrawal and transition to ongoing treatment.

- Rehabilitation/Residential – hospital (other than detoxification):

24 hours per day medical care in a hospital facility in conjunction with treatment services for alcohol and other drug use and dependency.

- Rehabilitation/Residential – short term (30 days or fewer):

Typically, 30 days or fewer of non-acute care in a setting with treatment services for alcohol and other drug use and dependency.

- Rehabilitation/Residential – long term (more than 30 days):

Typically, more than 30 days of non-acute care in a setting with treatment services for alcohol and other drug use and dependency; may include transitional living arrangements such as halfway houses.

- Ambulatory - intensive outpatient:

At a minimum, treatment lasting two or more hours per day for 3 or more days per week.

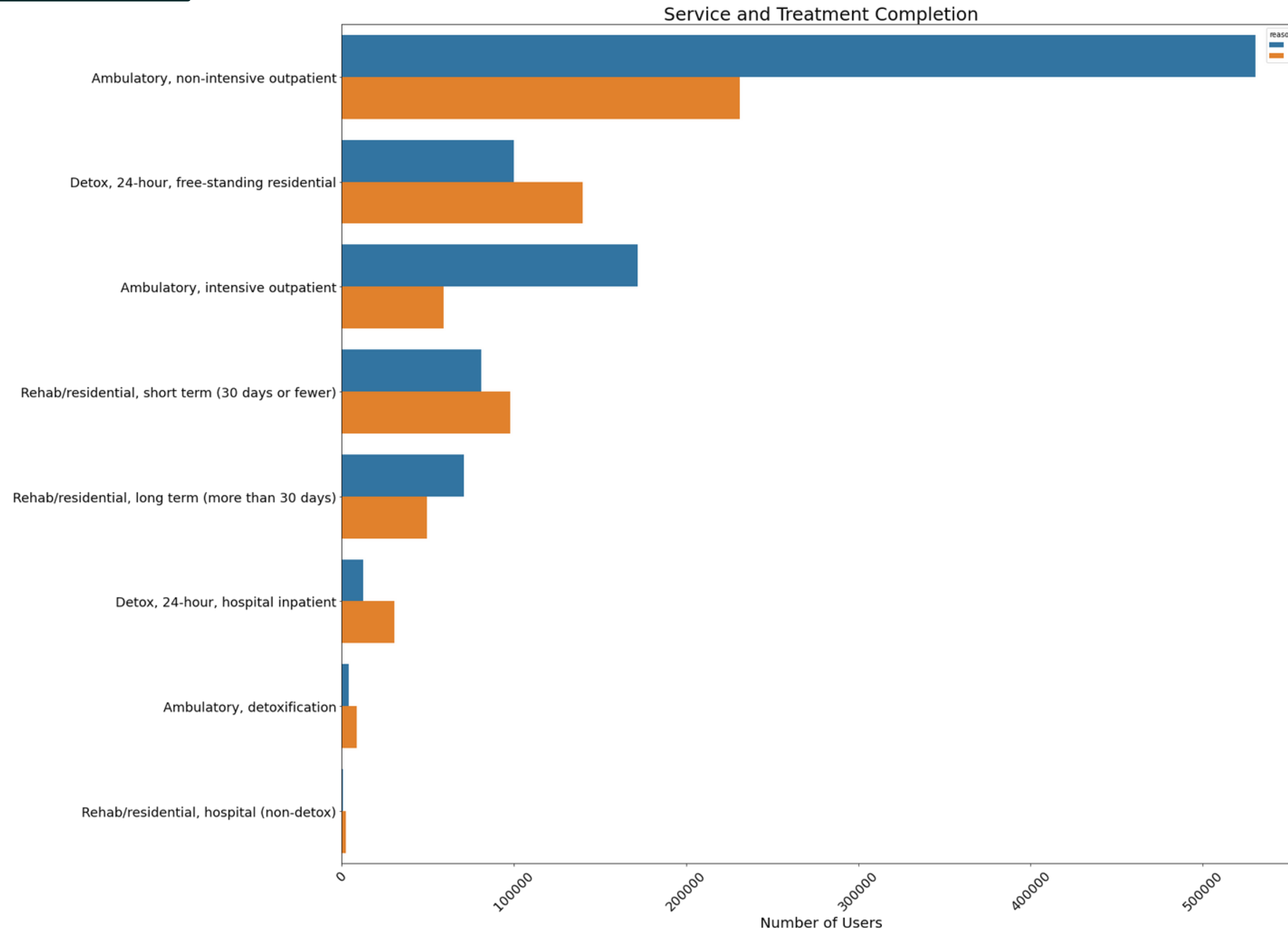
- Ambulatory - non-intensive outpatient:

Ambulatory treatment services including individual, family and/or group services; may include pharmacological therapies.

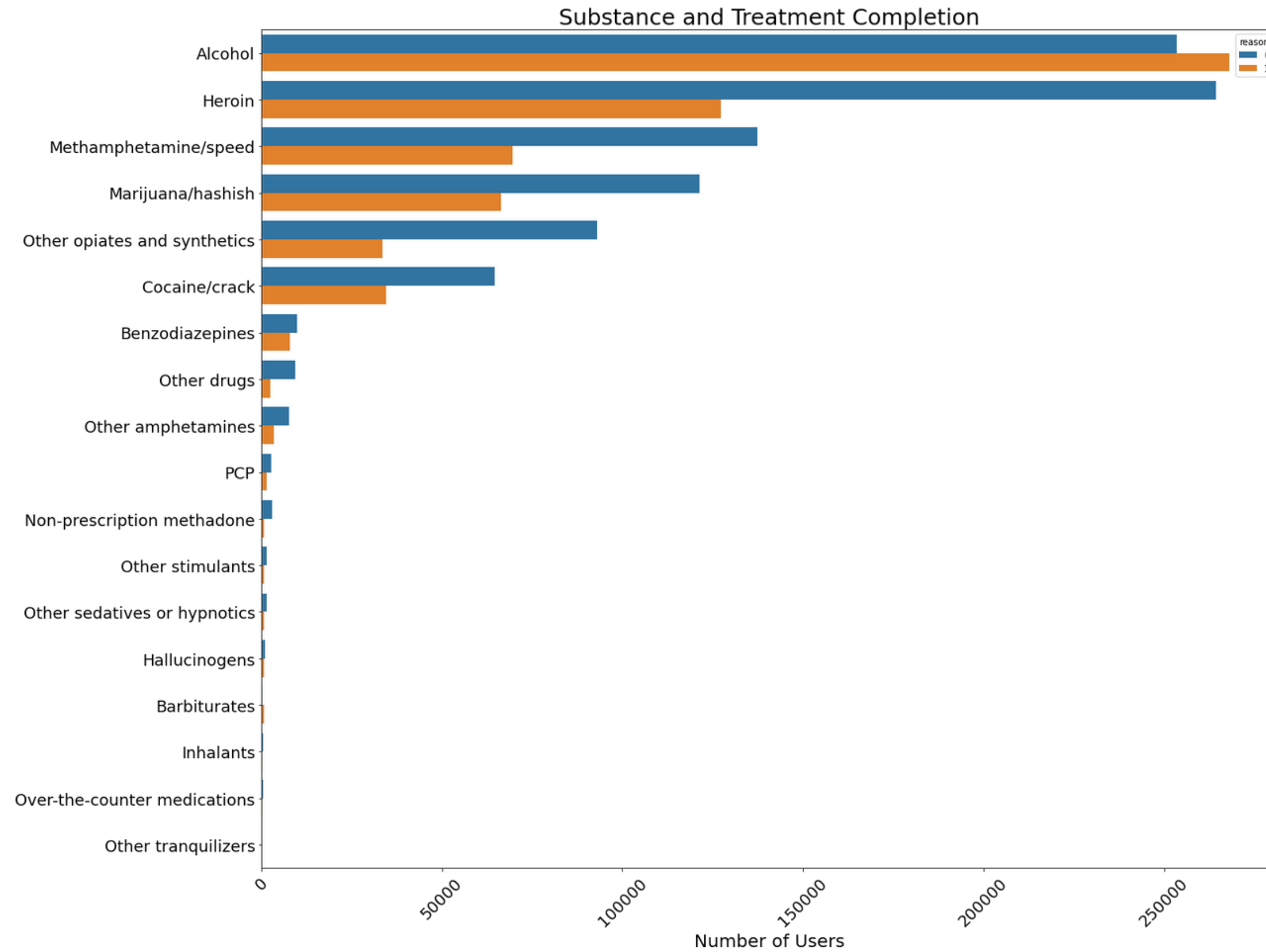
- Ambulatory - detoxification:

Outpatient treatment services providing for safe withdrawal in an ambulatory setting (pharmacological or non-pharmacological).

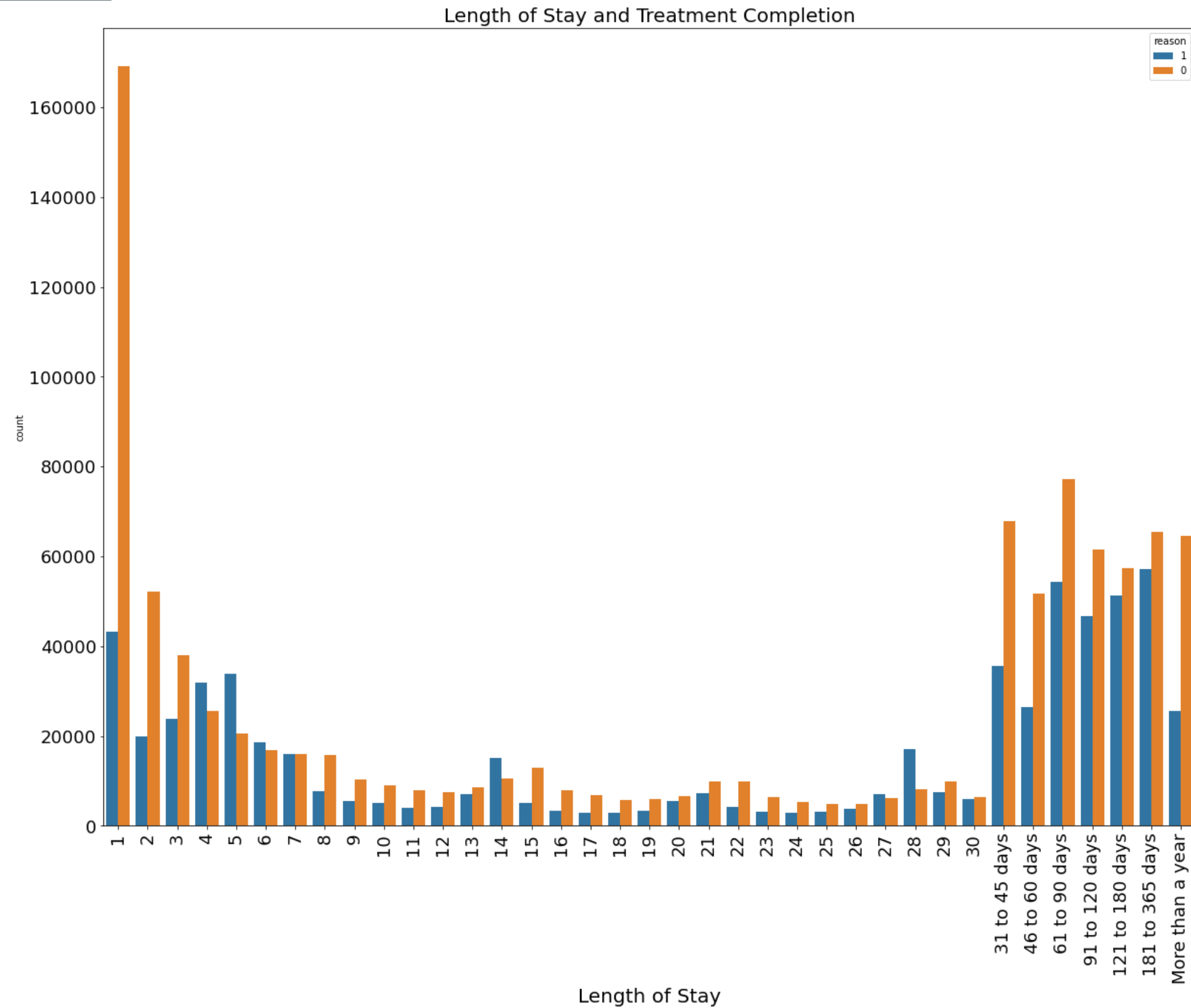
EDA



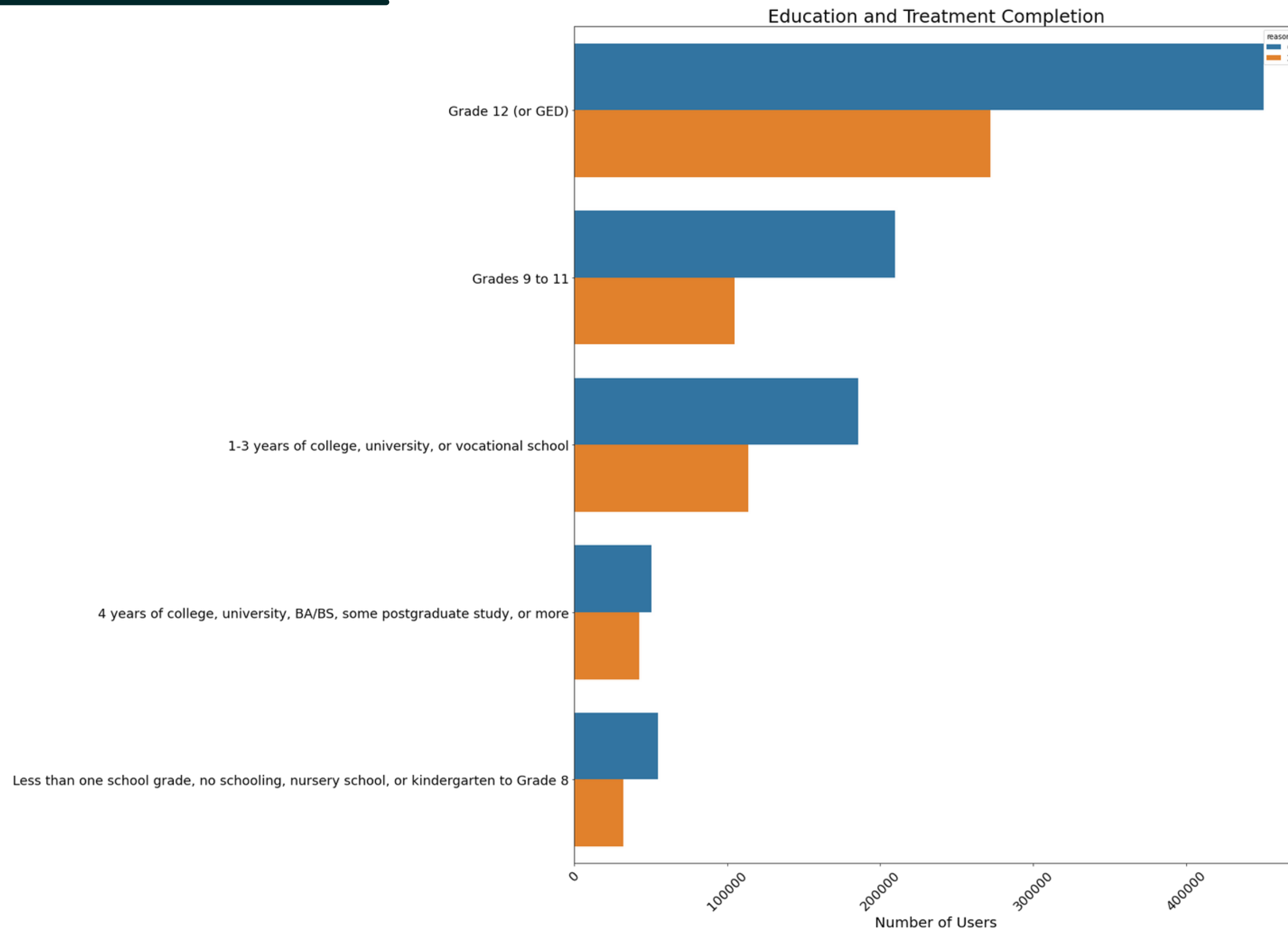
EDA



EDA



EDA



Feature Selection

- Chi2
 - Null Hypothesis
 - No relationship with the target variable
- Domain Knowledge
- Too many
- Relevance
 - Removed discharged applicable columns

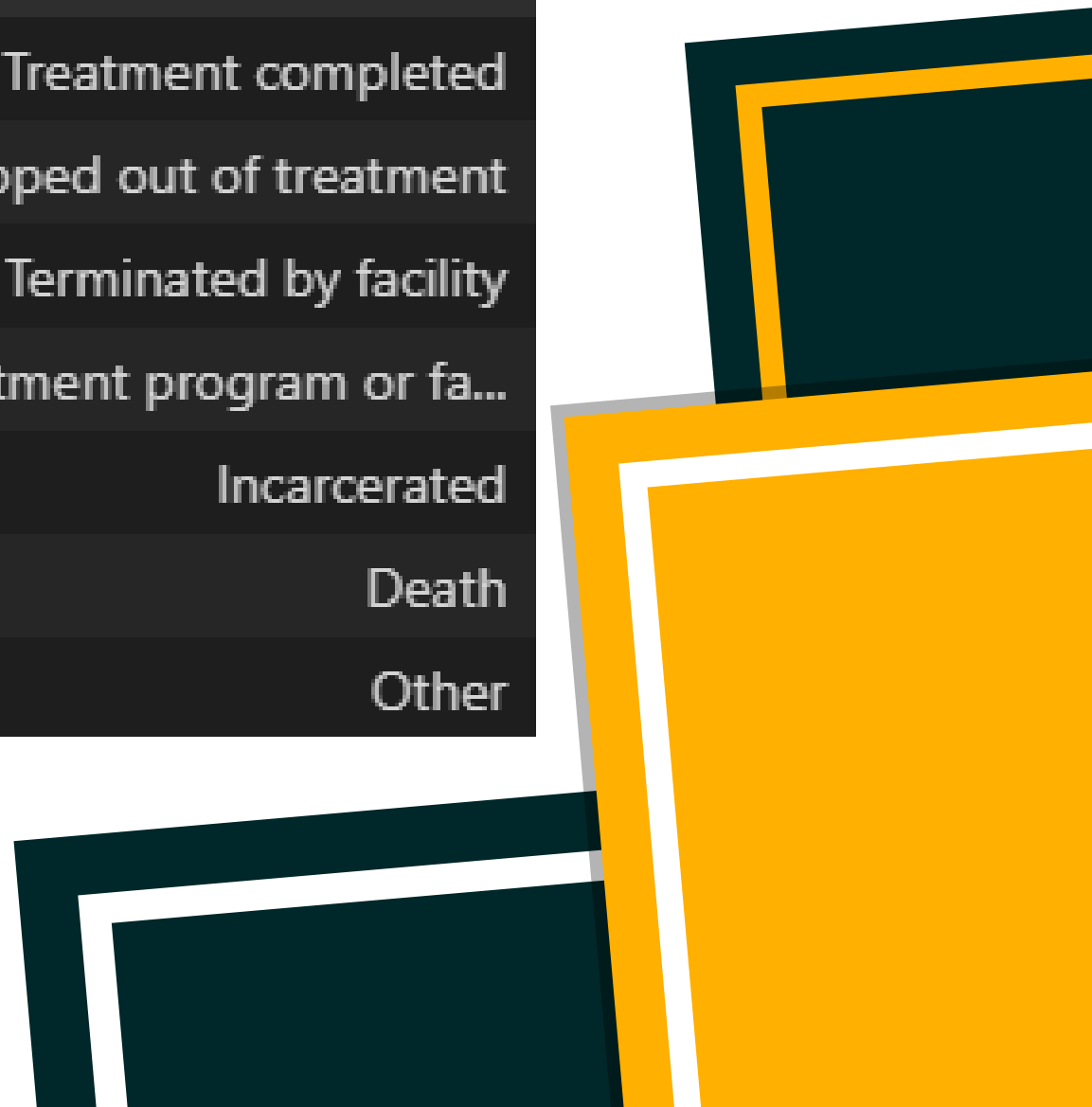
	features	crit_val	chi2	p_val	dof	rej_hyp
0	education	9.487729	5193.166586	0.000000e+00	4	True
1	marital_status	7.814728	471.832963	6.060739e-102	3	True
2	service	14.067140	122631.582765	0.000000e+00	7	True
3	len_stay	50.998460	89817.205442	0.000000e+00	36	True
4	ref_source	12.591587	8297.034337	0.000000e+00	6	True
5	treat_ep	3.841459	3649.349464	0.000000e+00	1	True
6	num_arrest	5.991465	2630.916663	0.000000e+00	2	True
7	empl_status	7.814728	22103.170202	0.000000e+00	3	True
8	gender	3.841459	6952.459628	0.000000e+00	1	True
9	housing	5.991465	4135.471654	0.000000e+00	2	True
10	diagnosis	28.869299	71195.589542	0.000000e+00	18	True
11	age_range	19.675138	10650.448876	0.000000e+00	11	True
12	p_income	9.487729	3500.362003	0.000000e+00	4	True
13	substance_1	27.587112	56539.603396	0.000000e+00	17	True
14	afu	12.591587	6344.792218	0.000000e+00	6	True
15	self_attend	9.487729	3984.028941	0.000000e+00	4	True

Feature Engineering



- Target Variable
 - Turn the target column (reason) into binary
- One Hot Encode categorical variables
- Balance out imbalanced classes
 - Oversampling

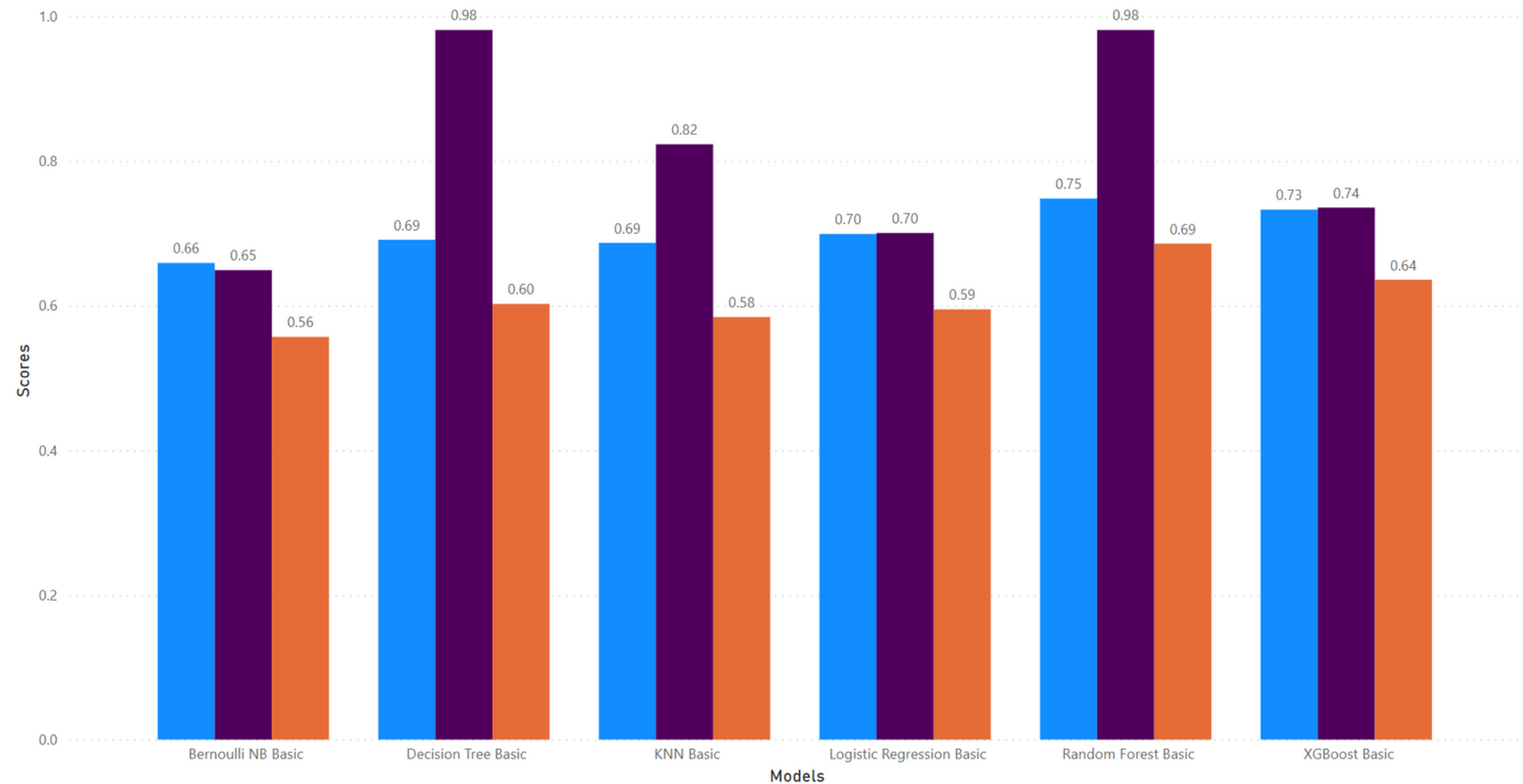
id	reason
1	Treatment completed
2	Dropped out of treatment
3	Terminated by facility
4	Transferred to another treatment program or fa...
5	Incarcerated
6	Death
7	Other



Model Performance

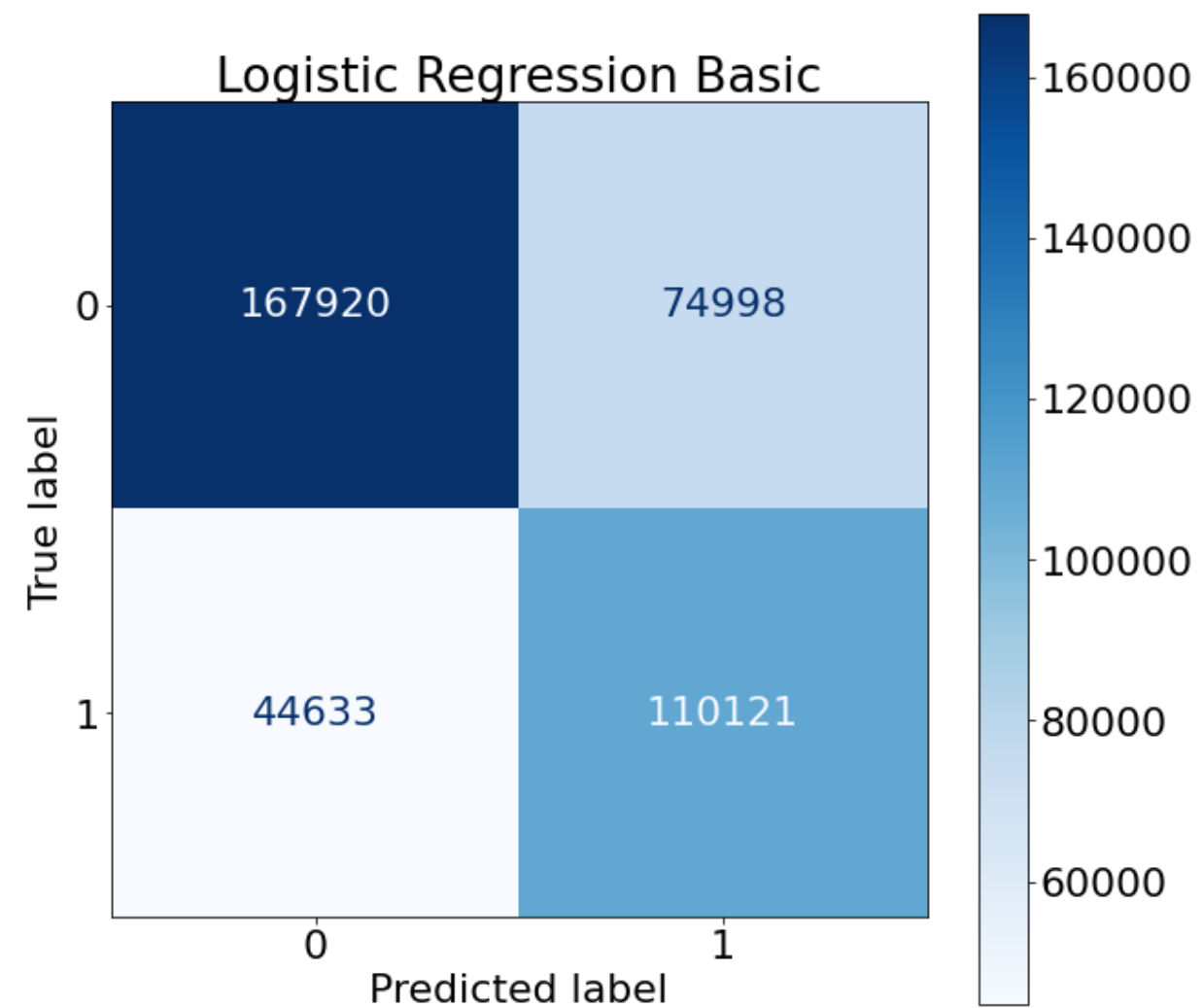
Basic Model Scores

● testing_score ● training_score ● precision

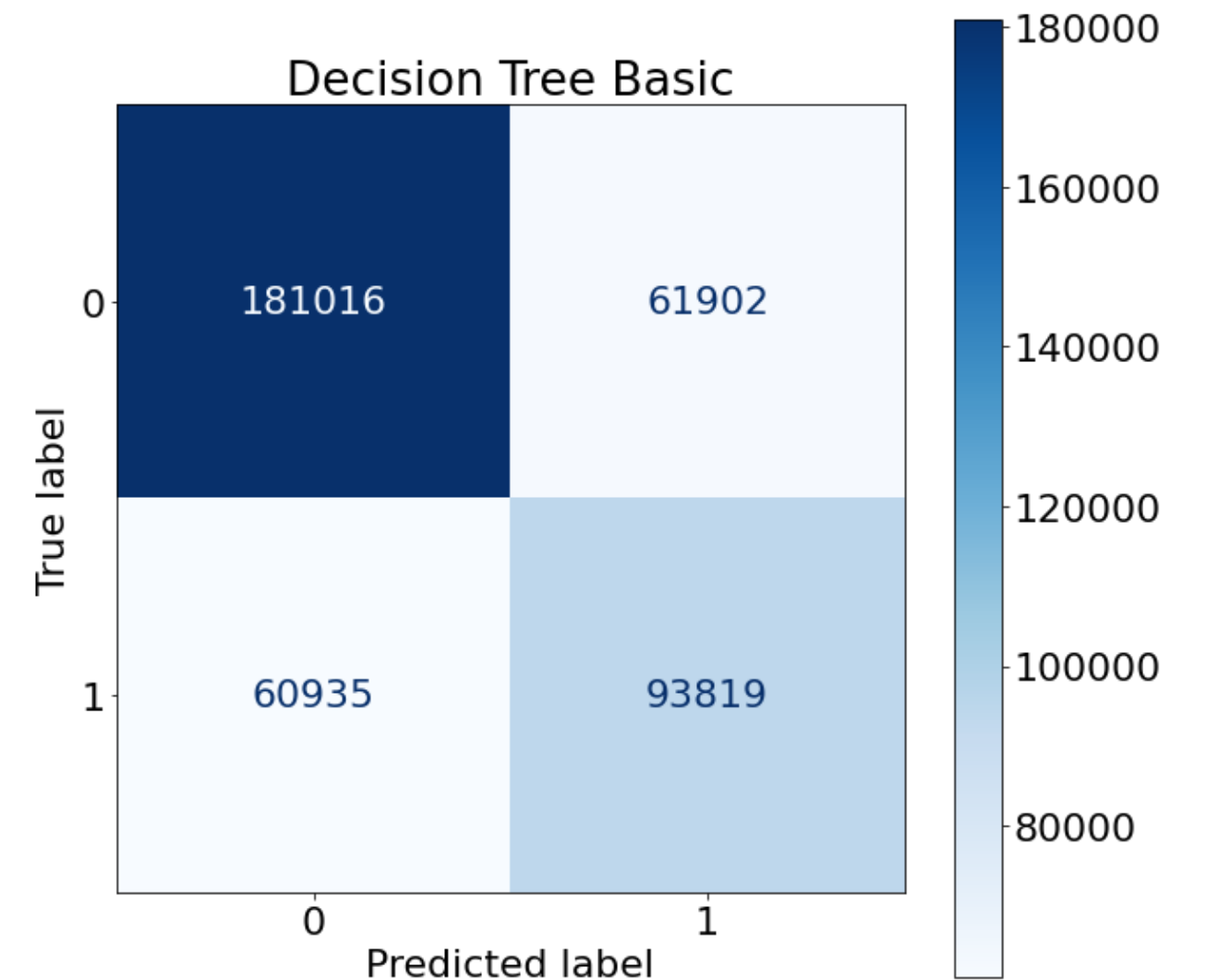


Models – Basic

01 Logistic Regression

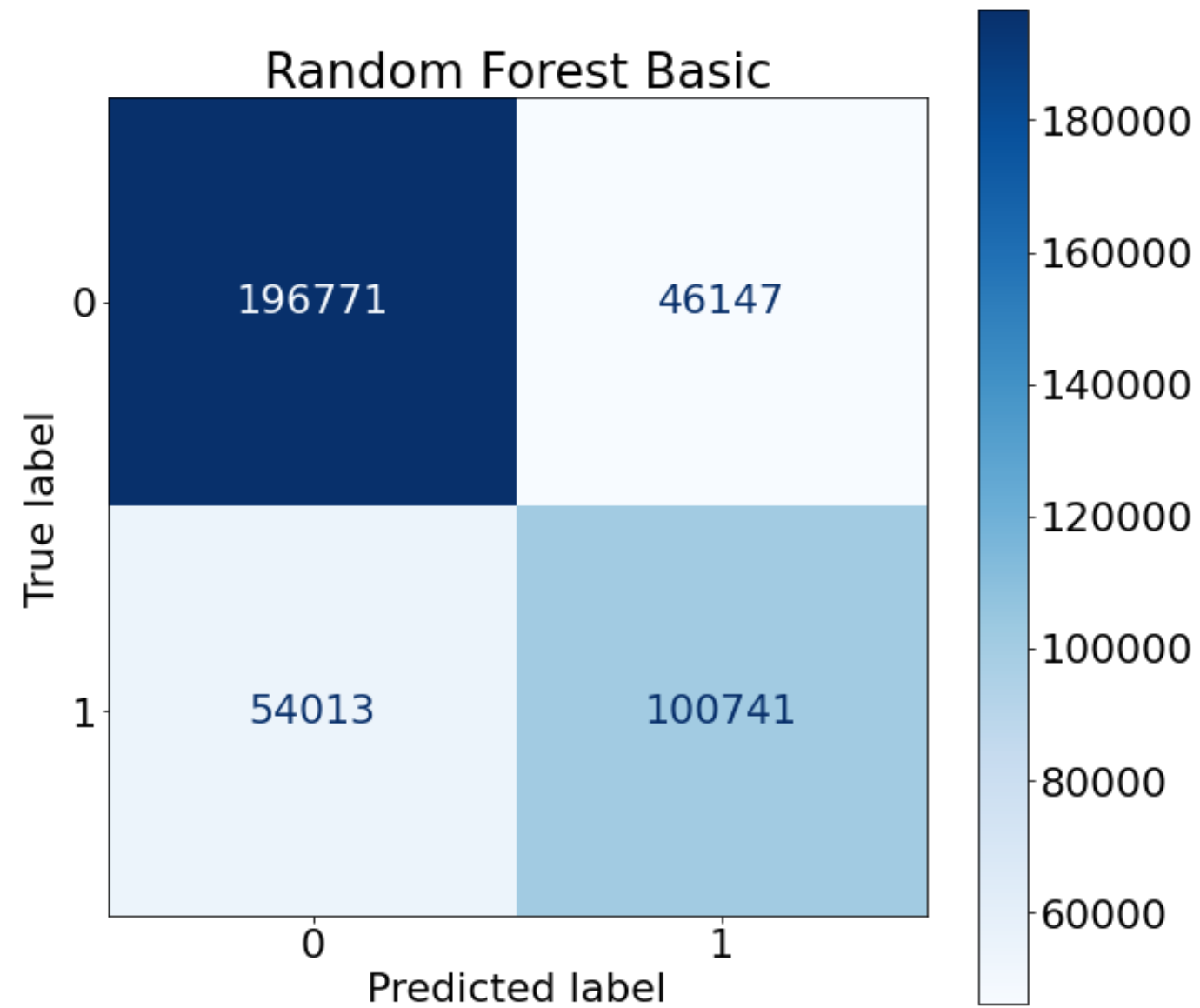


02 Decision Tree

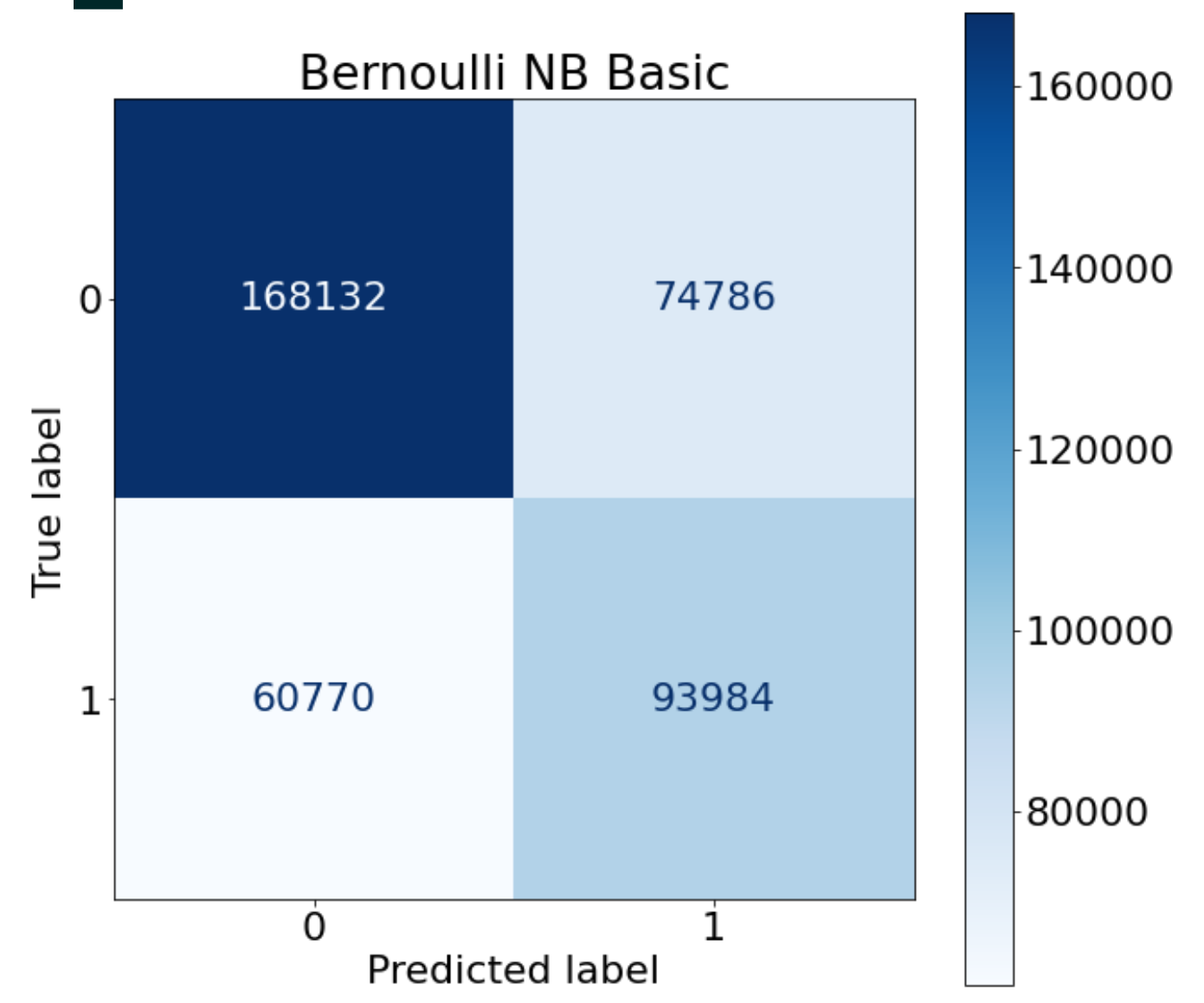


Models – Basic

03 Random Forest

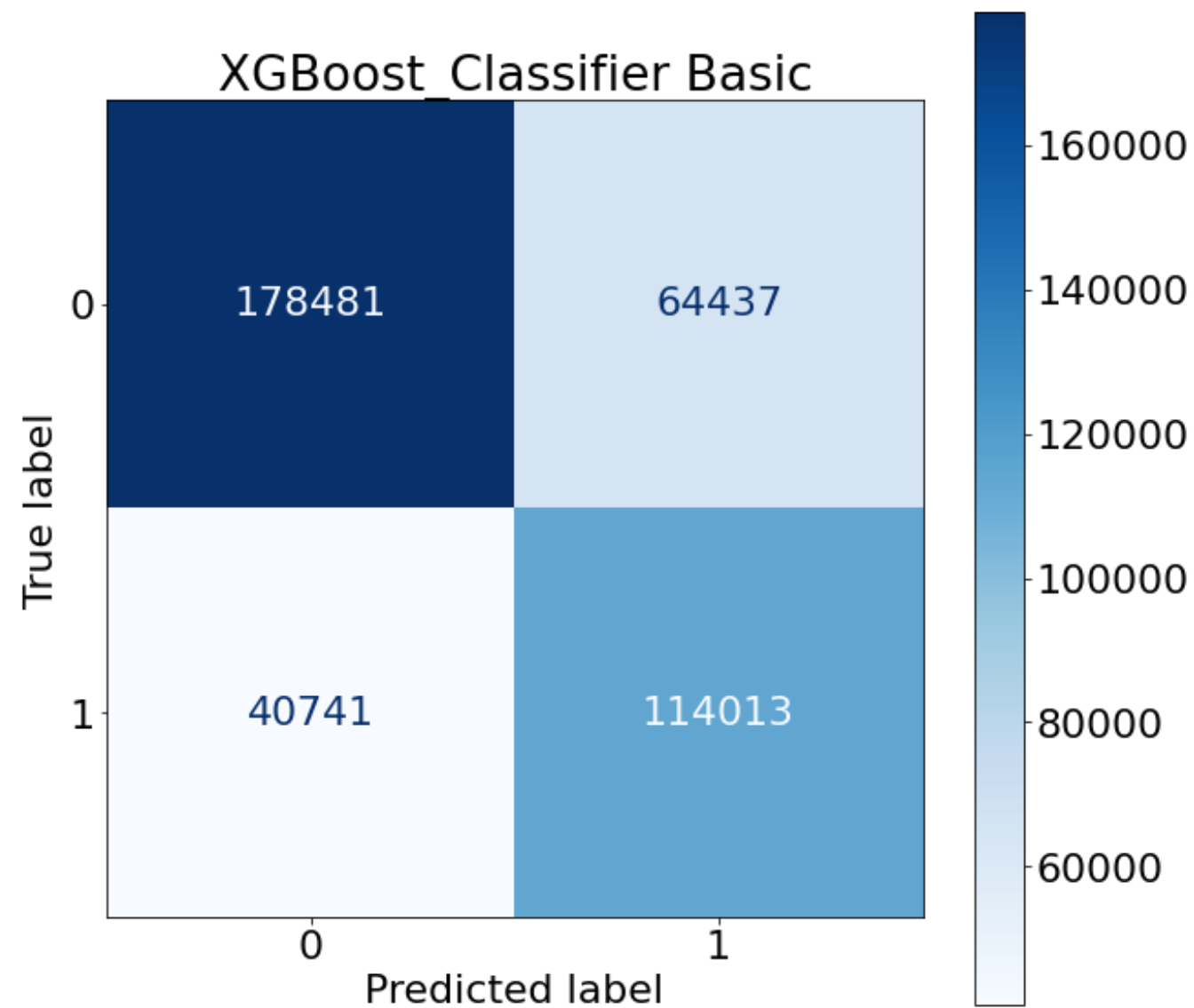


04 Bernoulli NB

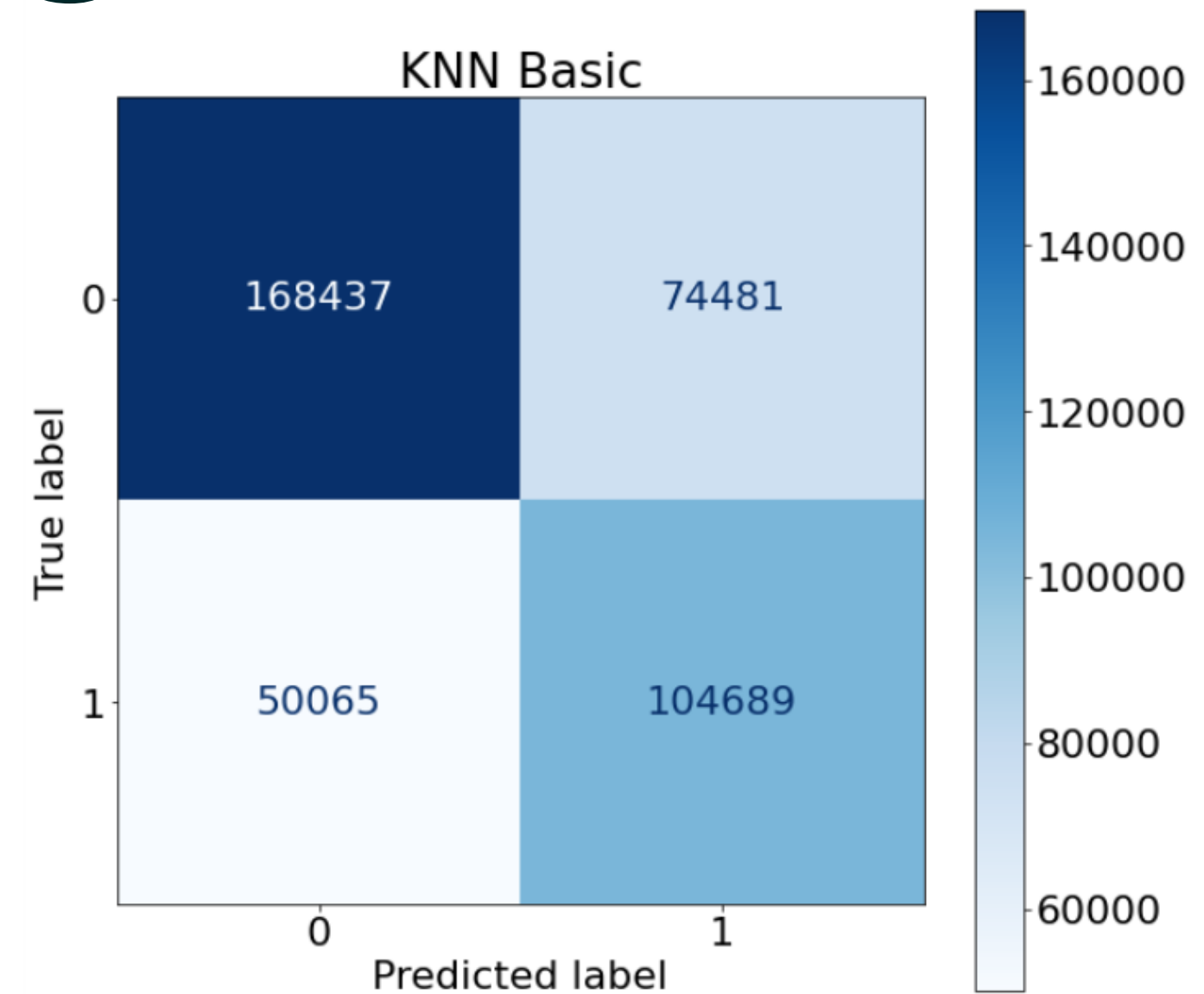


Models – Base

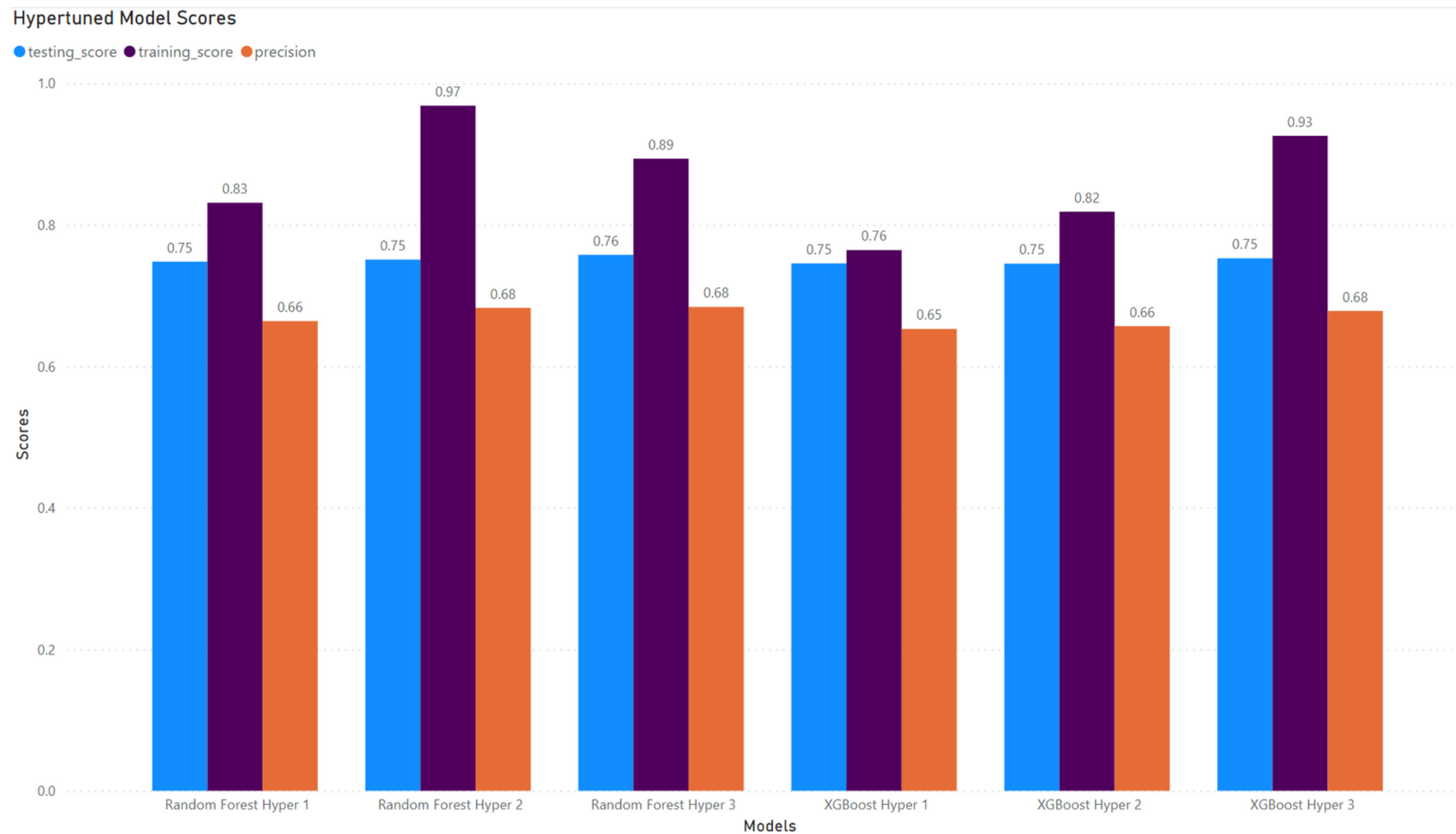
05 XGBoost



06 KNN

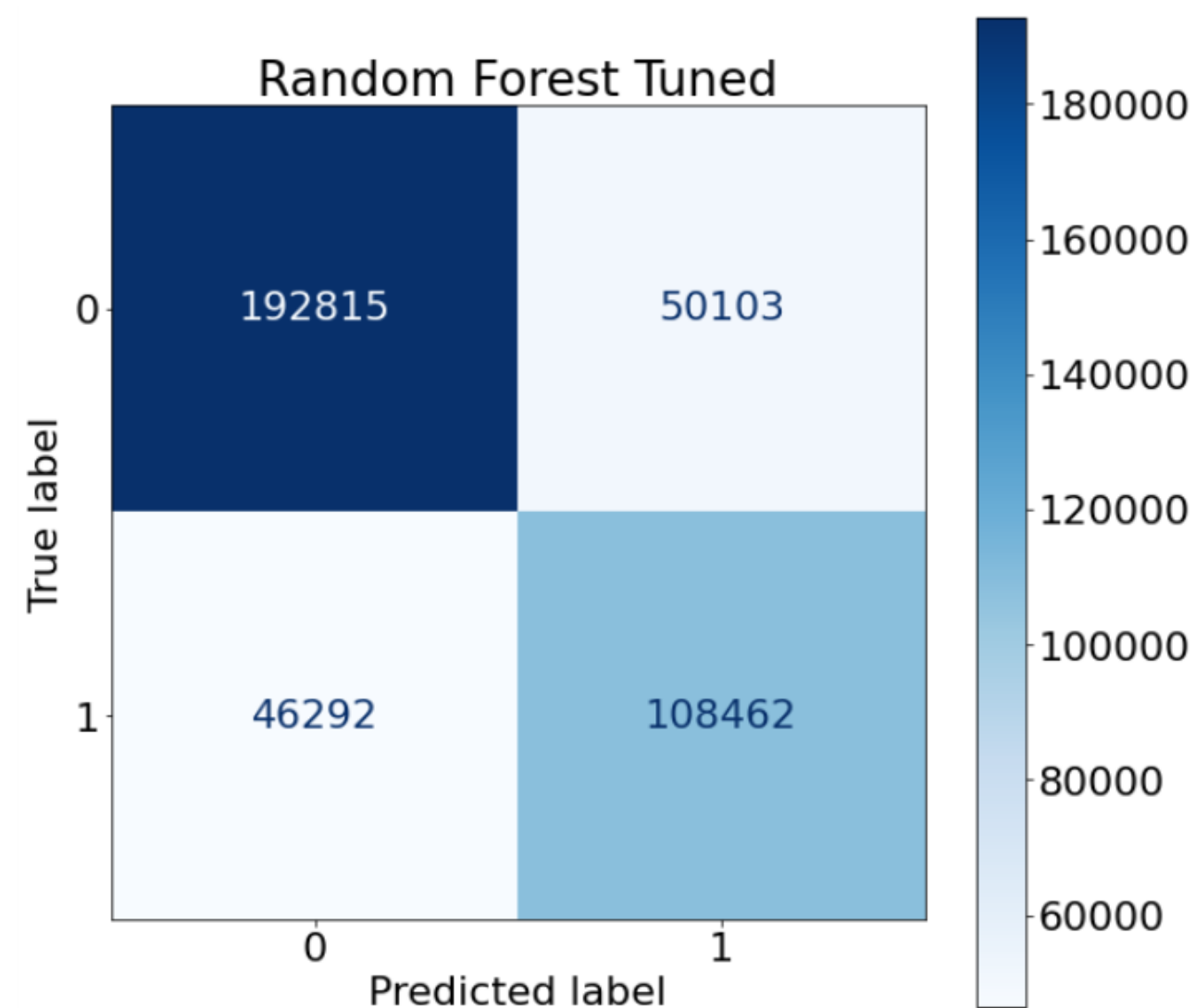


Model Performance

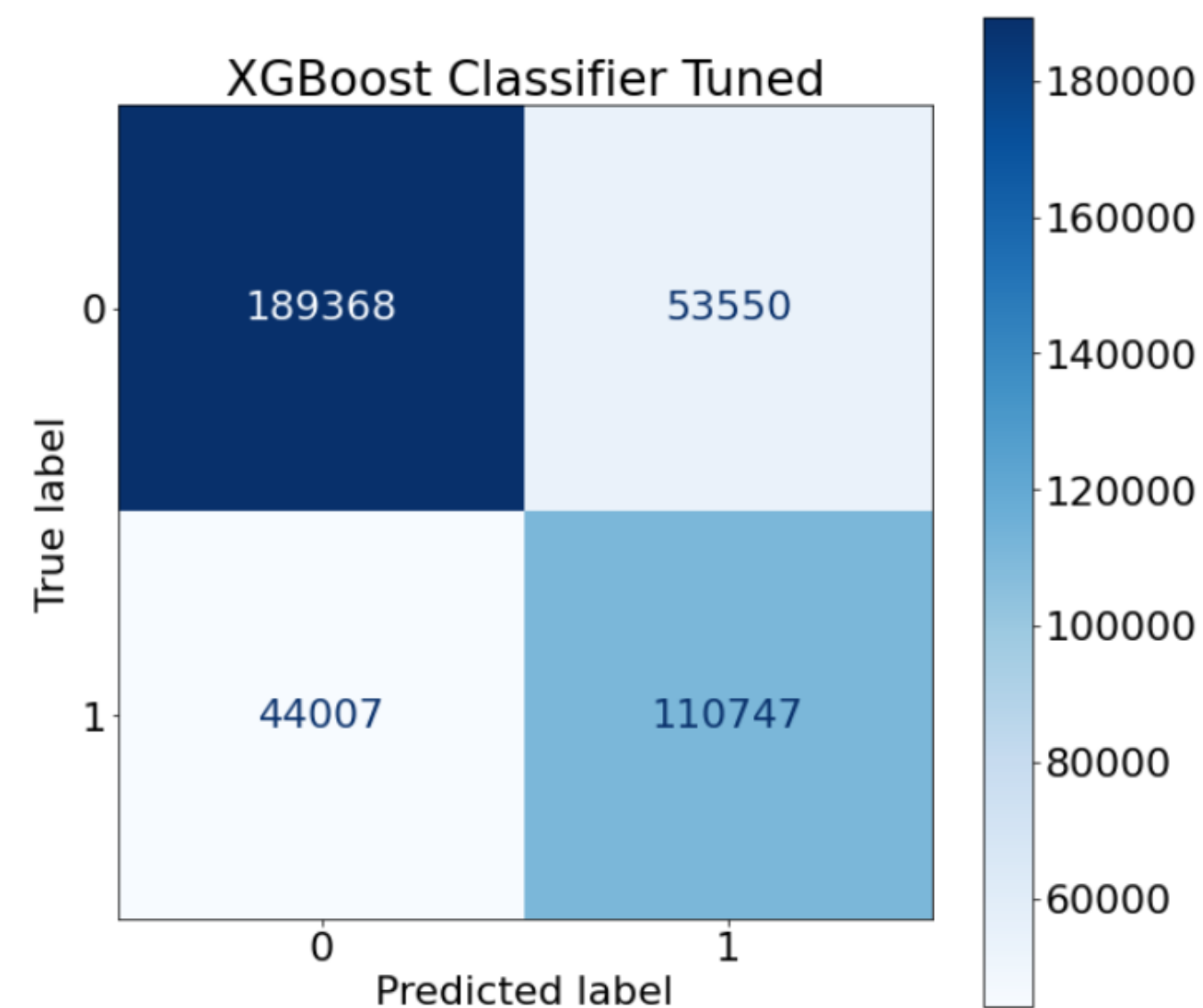


Models – Hypertuned

01 Random Forest

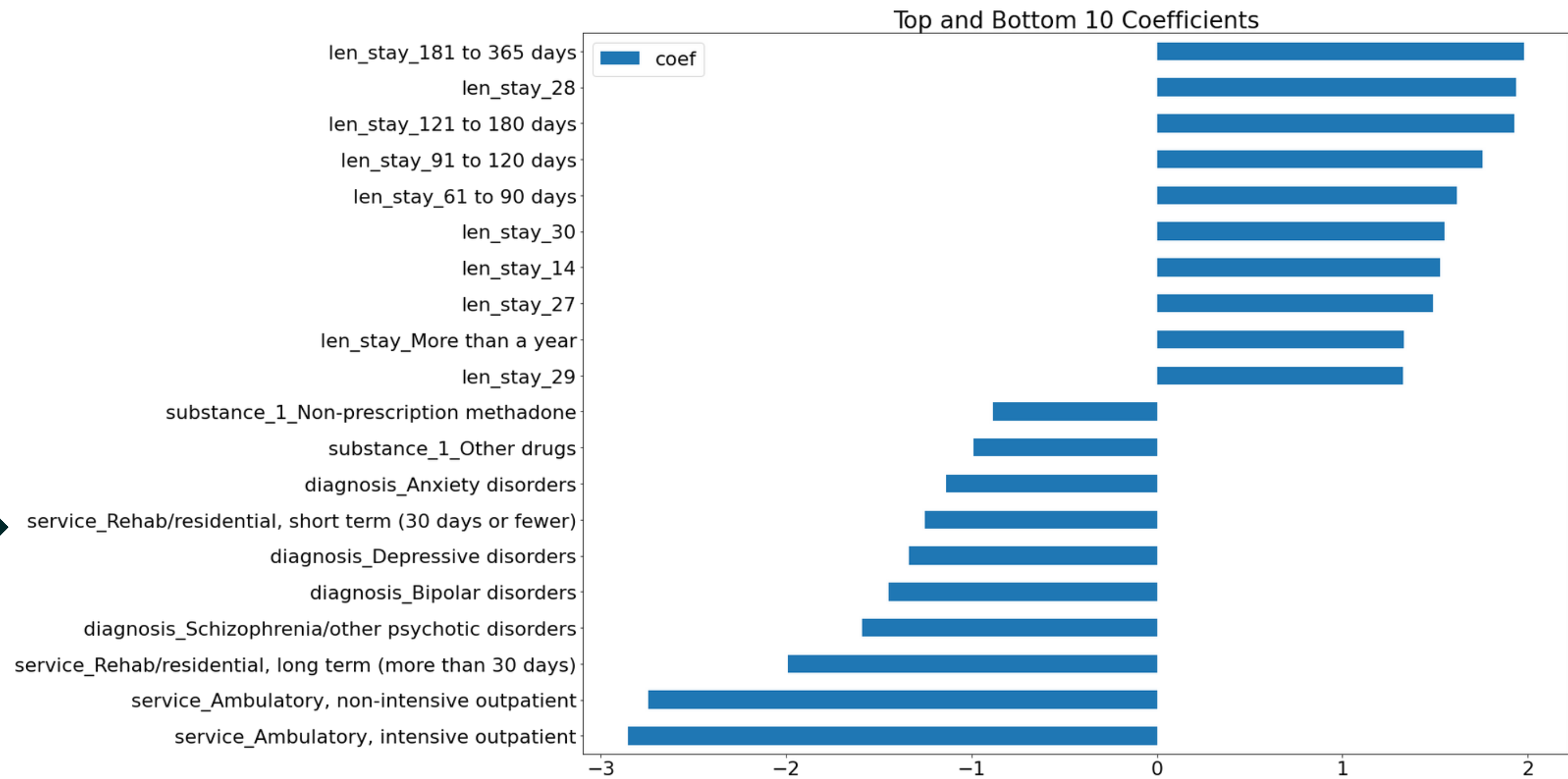


02 XGBoost



Findings

Feature Importance



Recommendations

Treatment Tool

- Use the model as a tool for treatment planning/case conferencing alongside assessment tools and the patient's own goals
- Try to reengage patients into treatment if they stop showing up or prematurely terminate service
- Feed updated information for better predictions

Next Steps

More Data

- 2018 and prior data
 - Retrain Model
 - Use for validation
 - Test Performance of the model of new data
 - Frequency of sessions
 - Individual Counseling
 - Groups
 - Overall time spent with a patient feature

Tailor Models

- Create models for different service types
 - Compare the performance of those models
- Multiclass classification

Next Steps

Preprocessing

- SMOTING or avoid using oversampling

Predictions

- Threshold

Further Feature Engineering/Extraction

- Experiment with categorical datatypes
- Combine Features together
 - Example: Substance 1 + 2 + 3
- PCA

Resources

- <https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm>
- <https://nymag.com/intelligencer/2021/03/deaths-of-despair-have-surged-among-people-of-color.html>
- <https://www.kaggle.com/code/hamelg/python-for-data-25-chi-squared-tests/notebook>
- <https://machinelearningmastery.com/feature-selection-with-categorical-data/>
- <https://github.com/scikit-learn/scikit-learn/discussions/20690>
- <https://stackoverflow.com/questions/61325314/how-to-change-plot-confusion-matrix-default-figure-size-in-sklearn-metrics-packa>
- <https://stackoverflow.com/questions/53574918/how-to-get-rid-of-white-lines-in-confusion-matrix>
- <https://stackoverflow.com/questions/67294768/how-can-i-change-the-font-size-in-this-confusion-matrix>
- <https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390>
- https://www.youtube.com/watch?v=ap2SS0-XPcE&t=987s&ab_channel=HarshKumar
- <https://towardsdatascience.com/xgboost-fine-tune-and-optimize-your-model-23d996fab663>
- <https://www.datafiles.samhsa.gov/dataset/teds-d-2019-ds0001-teds-d-2019-ds0001>

Public Service Announcement – XGBoost

- GridSearching
 - CPU only
 - Longer Training Time
 - CPU Temperature
 - GPU Support
 - Better Temperatures
 - Shorter Training Time
 - Conda Install
 - Only installs the CPU version
 - NVIDIA
 - Install Cuda

