Predicting San Francisco Fire Department unit availability time

based on historical data of calls for service data.

DSI Capstone Project by **Olga Iushchenko**

PROJECT IDEA

What about?

Project aims is to build a ML model that predicts time required for Fire Department unit (team) to solve the service and be available for next task. Model includes conditions available at the moment of call received by 911 and task dispatched to unit as well as historical trends.

Why is it important?

Correct distribution of resources increases effectiveness of Fire Department workflow and chances to provide in-time service where required. Having rich historical data and today’s technologies leads to smart usage of resources and planning tasks distribution within city.

DATASET PROFILE

4,7 M

datapoints in original dataset. updated weekly

3,4 M

valid datapoints used for model

34

features in original dataset

19

features in training dataset: 10 original and 9 engineered

April 2000 - July 2018

period of time reported

6

sequential chunks for cross validation

TOP FEATURES

Call Type

One of 27 categories for call. Top:

- medical
- fire
- alarms

Unit Type & ID

Type on engine/car used, one of 10 categories. Top:

- engine
- medic
- truck

Time of call

Includes hour and minutes of call received

Rolling mean

Rolling mean by types/units from past periods. Reflects trends for this group of calls

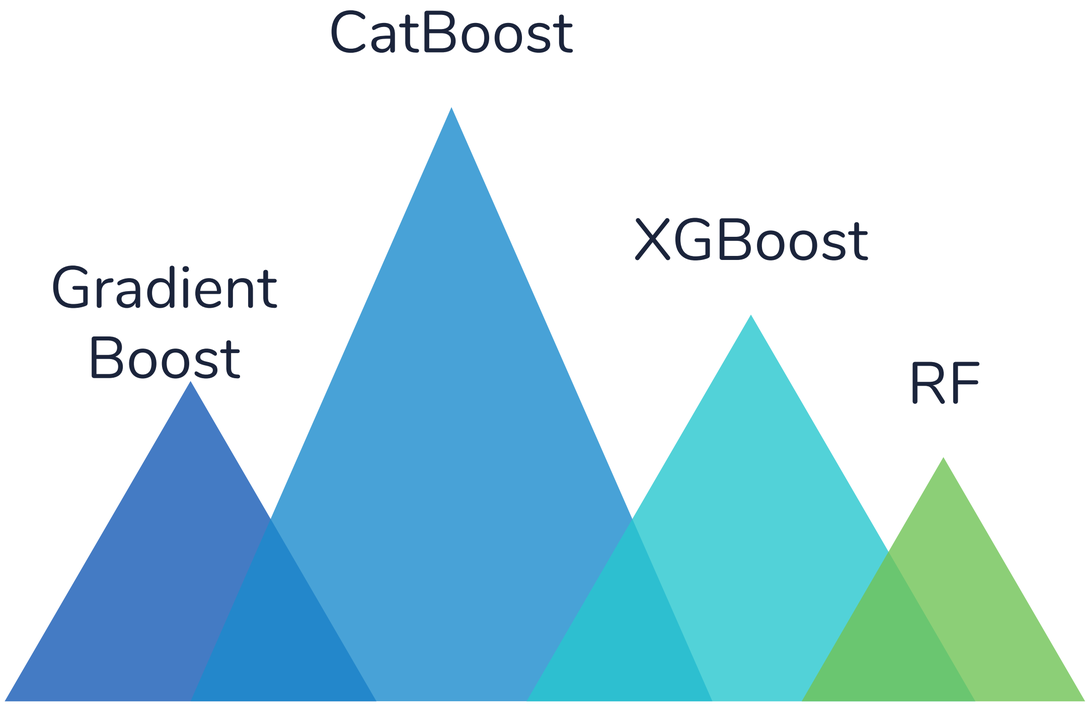
Zipcode & Box

Geographical features of call like Neighbourhood, Zipcode, Box, Coordinates

Original Priority

Original score set by 911 operator.

MODELS RESEARCH



CatBoost model choice:

- Problem required model to be flexible to multiple features
- Deep trees required for catching minor signal but effective interpretation
- CatBoost has convenient way of handling categorical features that are key for this problem

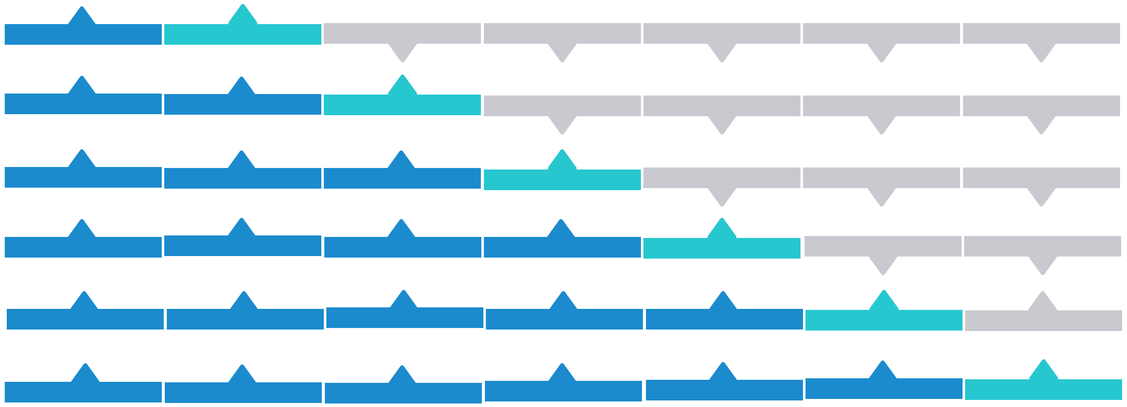
CROSS VALIDATION CHALLENGE

Using nested Cross Validation with time - series approach to avoid leakage from future records but validate model.

Mean evaluation scores are calculated as average score from each stage of validation.

*min

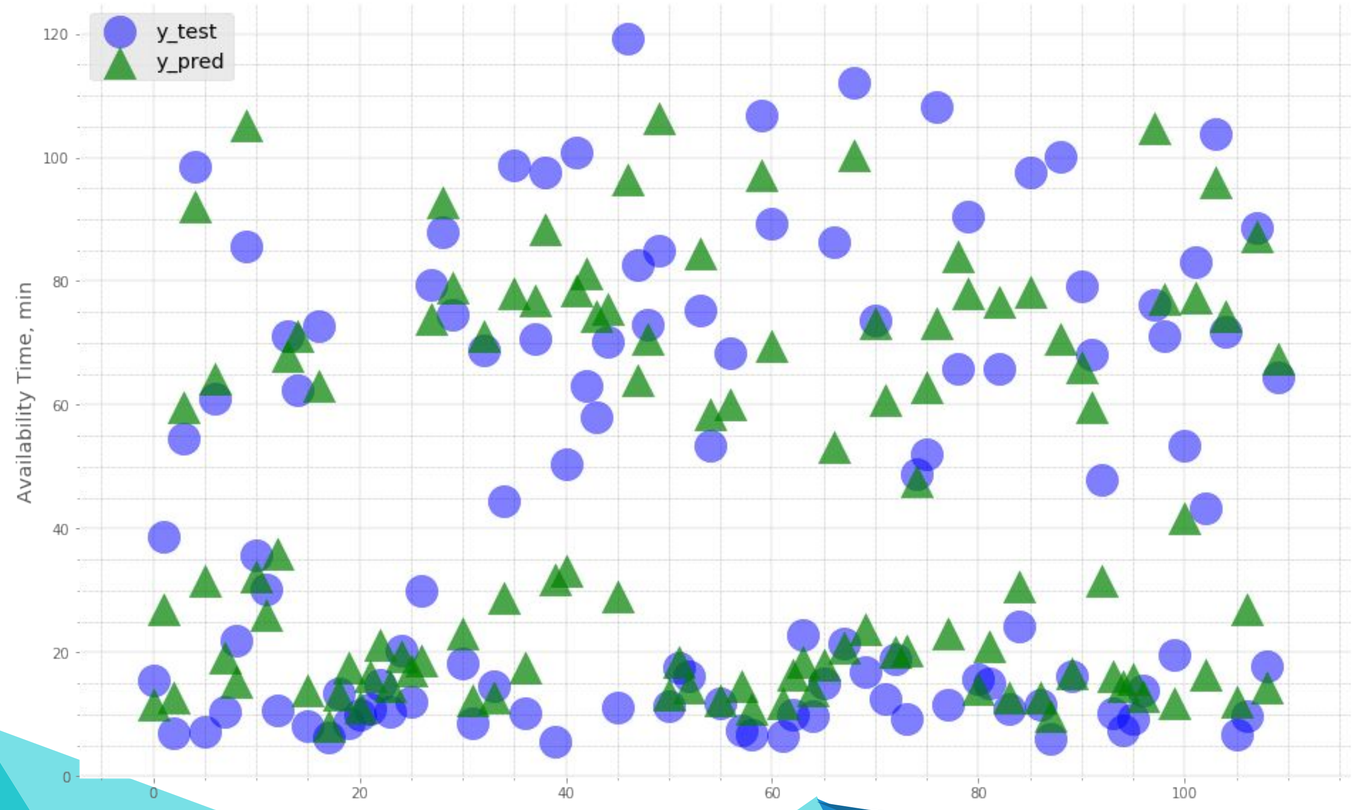
MAE	RMSE	R2 Score
9.73	14.25	0.77



Train Test

MODEL PERFORMANCE

Predicted vs Real Availability time (sample size 120 data points)



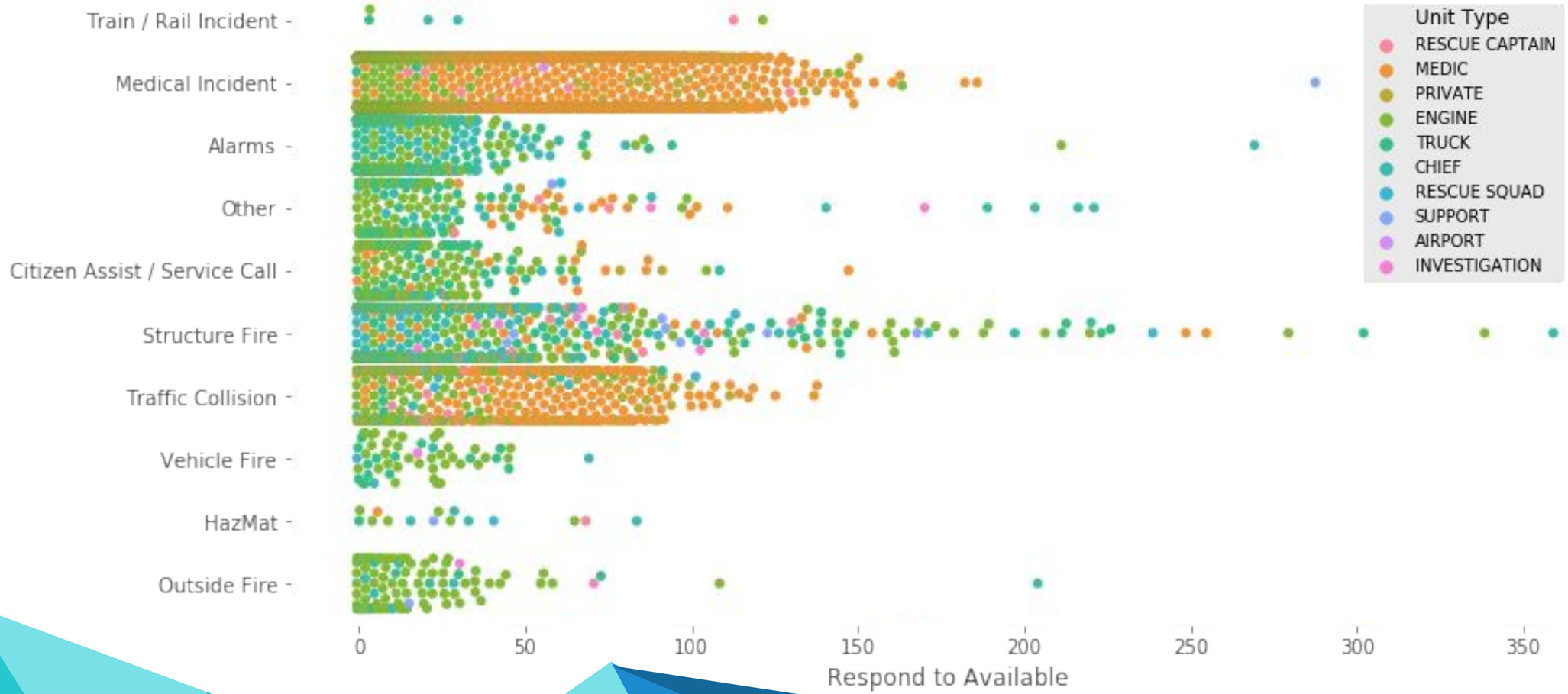
Model test scores:

MAE	RMSE	R2 Score
10.57	15.24	0.79

Dataset label mean = 34.75 min

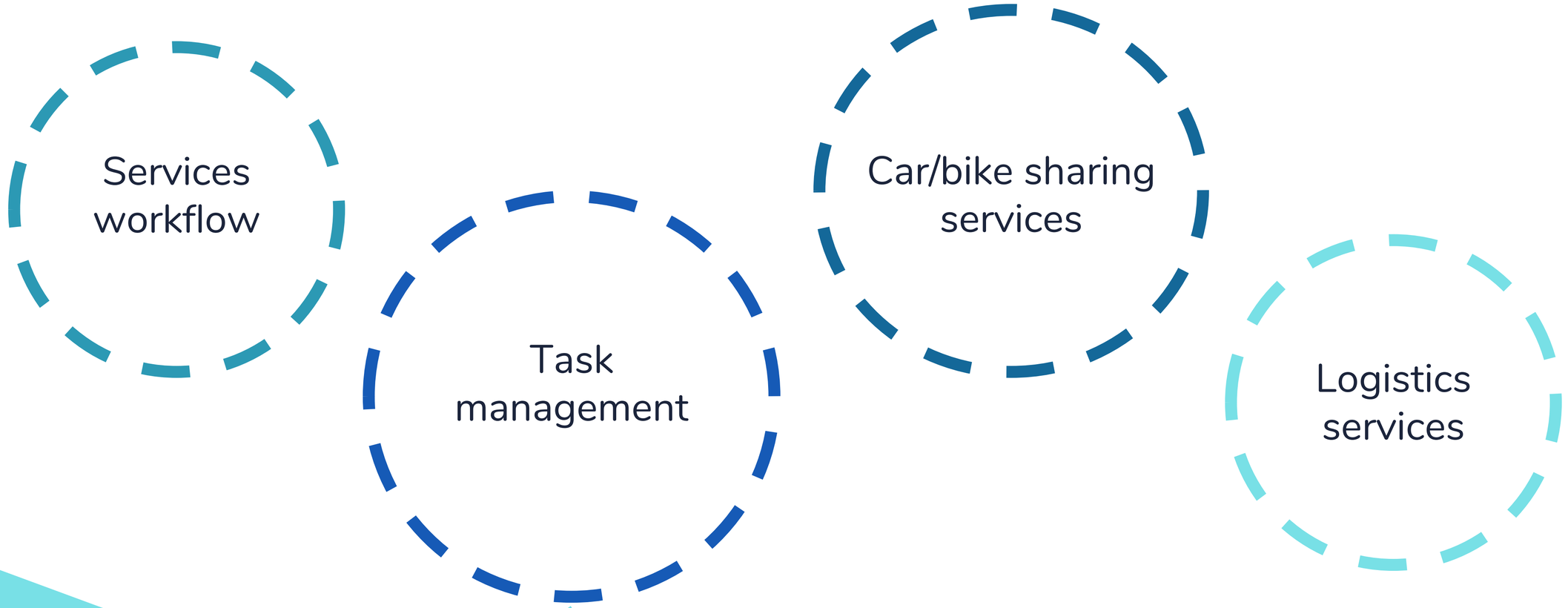
Within implementation model can be used for prediction multiple possible outcomes depending on new inputs to get better scores.

SAMPLE CATEGORIES TYPES



MODEL APPROACH VALUE

Although every problem requires individual approach to use all available data and find solution, this model approach may be used as base for range of problems:



Thanks!

Any questions?

You can find me at: /olga-iushchenko/
More about project: olgaiushchenko/ds-final-project