Luke Salamone

Simon Benigeri

Renpin Luo

Checkpoint 4: Machine Learning

For our machine learning checkpoint we looked at the following question:

> For officers with sustained allegations, can we predict the percent change in pay for an officer in a given year, given a number of variables such as whether the officer has had multiple previous sustained allegations, officer type, unit, race of officer, gender of officer, age of officer?

Prediction of Salary Raises for Police Officers

Approach 1: As a regression problem
Approach 2: As a time series problem

> **Script:** checkpoint_4_data.sql
> **Data:** checkpoint_4_data.csv
> **Notebook for approach 1 :** checkpoint_4_catboost.ipynb
> **Notebook for approach 2 :** checkpoint_4_time_series.ipynb

Context

Our group has been investigating the relationship between officer allegations and rewards within the Chicago Police Department, and has previously found troubling evidence that the two appear to be positively correlated. For checkpoint 1, we showed that in many years, officers with up to 5 allegations outperformed officers with no allegations in a given year. In checkpoint 2, we expanded upon this finding by illustrating that units with the most awards were the most frequently decorated. In checkpoint 3 our group illustrated a similarly troubling relationship between pay and sustained allegations, that as the number of allegations increases, the mean pay also increases.

The data

We queried the CPD database for officer ids and their respective salaries for each year in the period 2007-2015. Again, we chose this period because it is the one for which the dataset is the most complete.

We manipulated the data differently to suit each approach:

Approach 1: As a regression problem
Train Test Split Strategy:
TRAIN, Test = 80, 20
Train, Validation = 80, 20 (of Train)
DateTime features: year
Categorical features: race, gender
Features: salary, trr count, honorable mention count, allegations count, sustained allegations count, experience (year - year joined), age (year - date of birth)

Approach 2: As a regression problem for time series data
Train Test Split Strategy:
Index data by year
TRAIN, Test = Data where Year < 2014, Data where Year = 2014
Train, Validation = Data where Year < 2013, Data where Year = 2013

Categorical features: race, gender
Features: salary, trr count, honorable mention count, allegations count, sustained allegations count, experience (year - year joined), age (year - date of birth)

Methodology

Our machine learning calculations for this checkpoint can be found in the included python notebook.

Gradient boosting was an attractive choice for modeling our data because as an ensemble of decision trees it is more powerful than any one model alone. We used the CatBoost gradient boosting library for our regression analysis, which promises better default performance than other gradient boosting libraries. CatBoosts implements ordered boosting. Ordered boosting is described by the CatBoost team as “a permutation-driven alternative to the classic boosting algorithm.”

CatBoost’s algorithm is less prone to overfitting than those of Gradient Boosting libraries, like LightGBM and XGBoost. CatBoost’s ordered boosting creates what are called “oblivious decision trees” by using the same splitting criterion across an entire level of the tree. The resulting trees are more symmetric and less prone to overfitting, than those constructed by XGBoost and LightGBM.

In addition, CatBoost handles categorical variables out of the box. Sometimes we found that we needed to drop columns from one-hot encoded categorical features. In the case of gender, which, in this dataset, is binary, the preprocessing strategy is trivial: we can drop a column and retain all information. With race, it was more difficult. With CatBoost, we only

needed to pass the list of the categorical feature column names ‘race’ and ‘gender’ as a parameter, and we obtained better results than our initial experiments with one hot encoding for other algorithms.

Catboost also comes with functions for interpretability. For example, we used feature importance to analyze the results of our experiments.

We use the same algorithm for both approaches. We experimented with sklearn’s LinearRegression, MLPRegressor, KNNRegressor, and RandomForestRegressor. CatBoost yielded better results.

Analysis

For Approach 1: As a regression problem

		Root Mean Squared Error	Explained Variance	Max Error	Mean Absolute Error	Median Absolute Error	R2 Score
train	0	0.8915	0.9740	42.3957	0.1972	0.0522	0.9740
validation	0	1.0558	0.9641	34.4040	0.2242	0.0525	0.9641
test	0	1.0611	0.9601	36.9216	0.2154	0.0511	0.9600

Our regression analysis resulted in an R2 score of 0.96 on the blind test set. So ~96% of the variation in next year’s raise is accounted for by its regression on the selected features.

		Root Mean Squared Error	Explained Variance	Max Error	Mean Absolute Error	Median Absolute Error	R2 Score
train	0	2.421066	0.804596	46.247983	1.768394	1.687703	0.804594
validation	0	2.509589	0.792046	30.199449	1.846884	1.783552	0.791993
test	0	2.537193	0.790446	31.453095	1.832386	1.755273	0.790446

The feature importance can shed quite a bit of light on the effect of each of the measured factors on how much of a raise an officer received:

The feature importance metric tells us how much salience each of the given features had when modeling the regressor. Numbers are normalized to add to 100.

	year	salary	race	gender	trr_count	hm_count	allegations_count	sustained_count	experience	age
0	40.583177	46.517113	0.209922	0.137351	0.023156	0.340659	0.070062	0.011143	11.598437	0.508979

It makes intuitive sense that year, salary, and experience play the largest role in determining the salary increases for officers in a given year. Also, it is confirmed by our correlation analysis below:

	officer_id	salary	trr_count	hm_count	allegations_count	sustained_count	experience	age	raise
officer_id	1.000000	-0.002583	0.028239	0.030924	0.038289	-0.000480	-0.025091	-0.038374	-0.010735
salary	-0.002583	1.000000	-0.103111	-0.116256	-0.105376	-0.006115	0.734985	0.594448	-0.627751
trr_count	0.028239	-0.103111	1.000000	0.353181	0.352012	0.042530	-0.261296	-0.279220	-0.010721
hm_count	0.030924	-0.116256	0.353181	1.000000	0.355241	0.006597	-0.267680	-0.305088	-0.010970
allegations_count	0.038289	-0.105376	0.352012	0.355241	1.000000	0.230412	-0.187117	-0.195534	-0.045327
sustained_count	-0.000480	-0.006115	0.042530	0.006597	0.230412	1.000000	-0.011898	0.000077	-0.014986
experience	-0.025091	0.734985	-0.261296	-0.267680	-0.187117	-0.011898	1.000000	0.814890	-0.344328
age	-0.038374	0.594448	-0.279220	-0.305088	-0.195534	0.000077	0.814890	1.000000	-0.284793
raise	-0.010735	-0.627751	-0.010721	-0.010970	-0.045327	-0.014986	-0.344328	-0.284793	1.000000

Experience is significant. In fact, we found that officers received large salary bumps (about 30%) after graduating from the police training academy.

Other features are less salient but still can give us insight into the results of salary decisions. For example, it seems interesting that sustained_count, i.e. the number of sustained allegations an officer had in a given year, was the least important feature, playing little to no role in deciding salaries in the next year. However, this may be explained by the fact that sustained allegations are a relatively uncommon occurrence. This is in contrast to hm_count, the number of honorable mention awards an officer received, which was much more salient.

For Approach 2: As a regression problem with Time Series data

		Root Mean Squared Error	Explained Variance	Max Error	Mean Absolute Error	Median Absolute Error	R2 Score
train	**0**	2.2509	0.7811	46.2116	1.4750	0.8457	0.7811
validation	**0**	5.8459	0.9399	28.5736	5.6555	5.6799	0.3394
test	**0**	1.6877	0.9118	28.4466	1.0556	0.6584	0.9068

We get high RMSE and low R2 scores on the validation set, probably because in 2013 the year police officers received the highest mean and median raises. Also, accounting for Year as we did in approach 1 seems to yield better predictions (Lower RMSE, higher R2 score). We see below that salary varies with year and some years, like 2013 and 2009 correspond to an exceptionally high mean raise. This explains the poorer results on the Validation Set, where Year=2013, but strong results on the blind, Test Set, where Year=2014.

	year	mean raise	median raise	std raise	var raise	skew raise
0	2007-01-01	2.264243	0.000000	5.183327	26.866884	3.570471
1	2008-01-01	1.610521	0.000000	3.340903	11.161635	4.741325
2	2009-01-01	8.115766	6.700016	3.028777	9.173489	4.751789
3	2010-01-01	3.517023	1.961730	3.987775	15.902352	5.413427
4	2011-01-01	1.908440	0.989202	2.529715	6.399456	7.040216
5	2012-01-01	1.978371	0.000000	6.218376	38.668205	4.046976
6	2013-01-01	8.261752	5.768952	7.192851	51.737108	3.347655
7	2014-01-01	3.123902	0.991678	5.527625	30.554638	4.475517

Again, we provide feature importance. Salary and experience are the most significant.

	salary	race	gender	trr_count	hm_count	allegations_count	sustained_count	experience	age
0	70.81794	0.329721	0.200052	0.230815	1.421175	0.383933	0.084362	24.709683	1.82232

Future Research

As far as the existing database is concerned, our team has tried as many methods and algorithms as possible to maximize the performance of our model, and has observed some valuable information in the process. The biggest obstacle we have encountered in this constant experimentation is the incompleteness of the dataset. The dataset is incomplete because we focus on a snapshot of the CPD from 2007 to 2015. To study the evolution of salaries, we may benefit from zooming out a bit more.

We could suggest that avenues for additional research in this area include diving deeper into some of our features, such as elements from the TRRs (firearm_used, subject_armed, subject_injured, etc). But we are weary of trying different sampling strategies until we see the results that we want. We lack domain knowledge of the inner workings of the police department payroll decisioning process. This may have led us to choose a subset of features which may be correlated but not causal of pay increases. Before we consider different strategies for feature selection and engineering, we should speak to domain experts.

We think that another useful avenue to explore is consideration of political and economic climate. The variations we see in raises from year to year may be explained by the world in which the CPD exists. For example, in 2008, we had the Wall St crash and the Obama election. In 2009 we saw a big year for police officer raises. In 2012 we had the reelection of President

Obama, and in 2013 we saw another big year for police officer raises. Perhaps these are linked, perhaps not. What we can assert is that we would benefit from a longer time period to look at in the data, and that we should consider significant events in Chicago and more broadly, the US. The motivation for these questions is simply the fact that the CPD is not isolated from society. We may see that a feature is salient, but our dataset does not describe the world at large. In this respect, using data from other police forces for the same years could be very useful.