EDA

&

CatBoost model

on cv_programmers dataset

made by:

Goryachev Alexander

27.07-01.08

Libraries:

datetime

AutoViz

Pandas

json

CatBoost

NumPy

requests

shapely

re

Optuna

mplcyberpunk

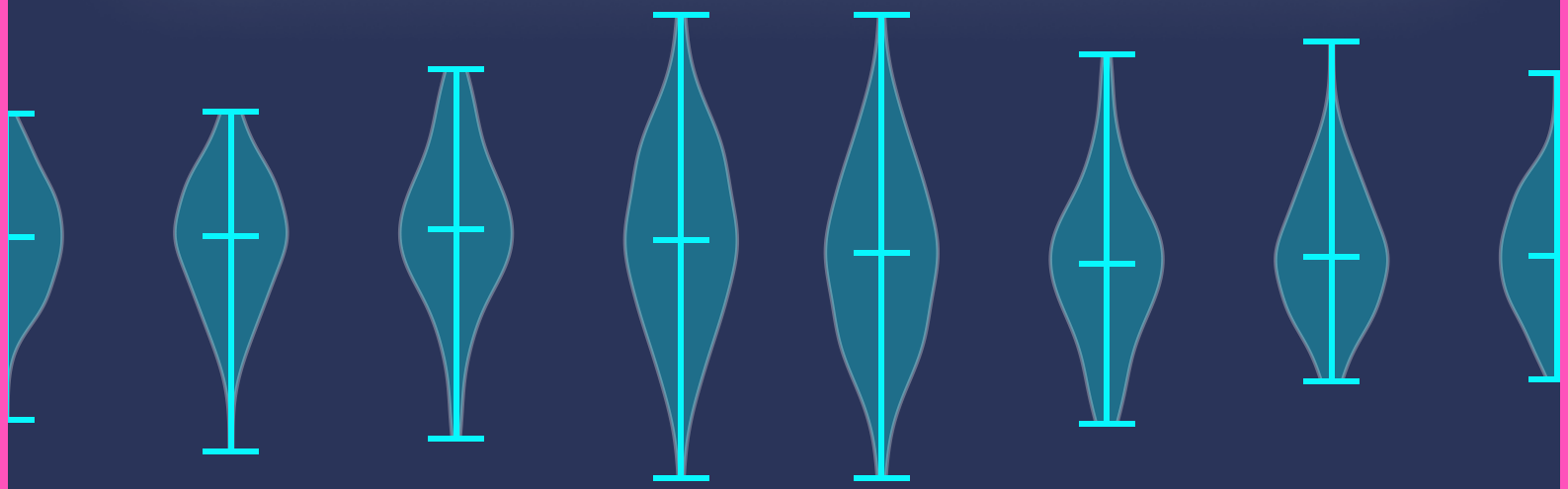
SciKit

Geopandas

SeaBorn

ast

Matplotlib



Plan

- Basic data cleaning
  - Dropping unnecessary columns
  - Encoding categorical variables
  - Filling NaN’s
  - Clipping anomalies
  - Adding new features
- Exploratory data analysis
- Building the model
- Evaluation

Data

Original dataset contains only 425 rows and 47 columns. Most of those columns are filled with NaN's, has zero variance or have categorical values which are all unique.

Therefore after cleaning the Dataframe we're left with 15 columns, which can help the model predict the target.

Those include:

- home_city_code
- education
- shedule_type
- experience
- professional_skills
- add_skills
- busy_type
- business_trip
- retraining_capability
- fullness_rate
- n_jobs
- additionalEducation
- english_level
- publication_date

Categorical to numeric

There were some categorical columns would be better to make numeric. CatBoost is good at handling categorics, but they nature suggest giving each value in them some weight.

For example:

EDUCATION(‘Среднее’ , ‘Среднее-профессиональное’, ‘Незаконченное высшее’, ‘Высшее’) -> EDUCATION(0.5, 0.6, 0.7, 1)

BUSY_TYPE(‘Полная занятость’, ‘Удаленная’, ‘Частичная занятость’, ‘Стажировка’, ‘Временная’) -> BUSY_TYPE(1, 0.9, 0.6, 0.5, 0.5)

Salary anomalies

Since the target variable is salary, thus making it it the most important one, we need to explore it thoroughly.

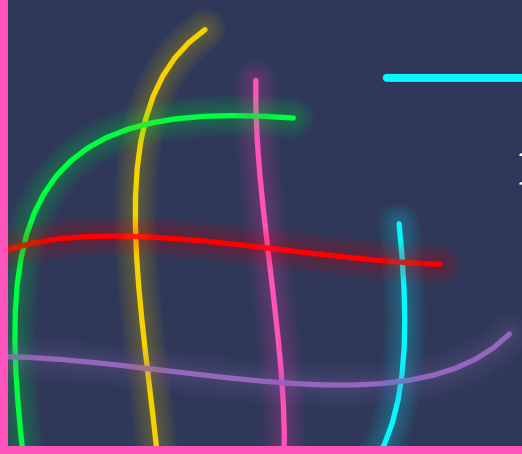
data.salary.describe() gives warning results:

mean	46426
std	40432
min	0
25%	25000
50%	40000
75%	50000
max	350000

Salary anomalies

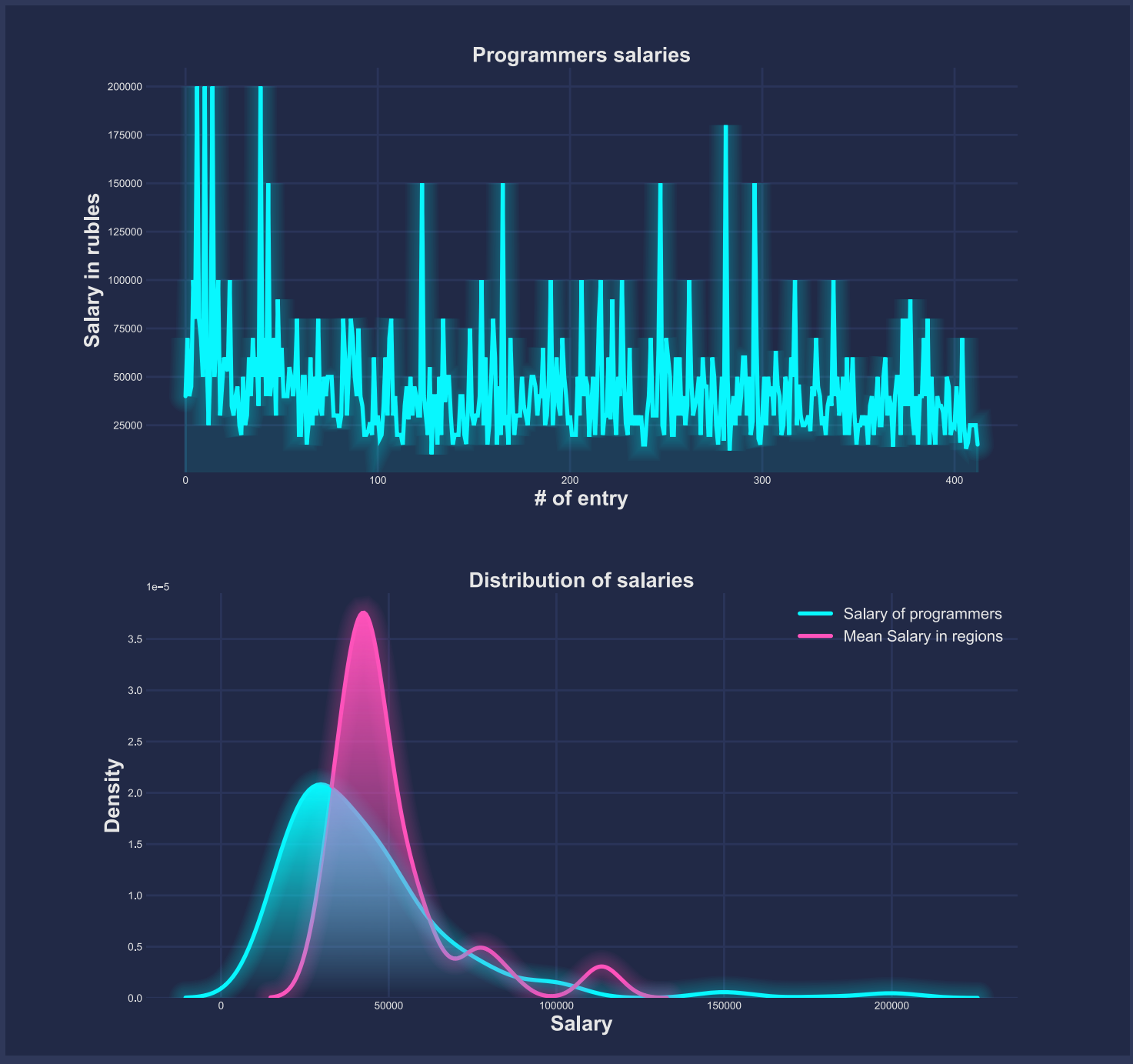
After capping top salary as 200k and deleting all the rows with salary less than 10k we have the following results:

mean	43626
std	28432
min	0
25%	25000
50%	40000
75%	50000
max	200000



Salary anomalies

That gives us plots like those:



New features

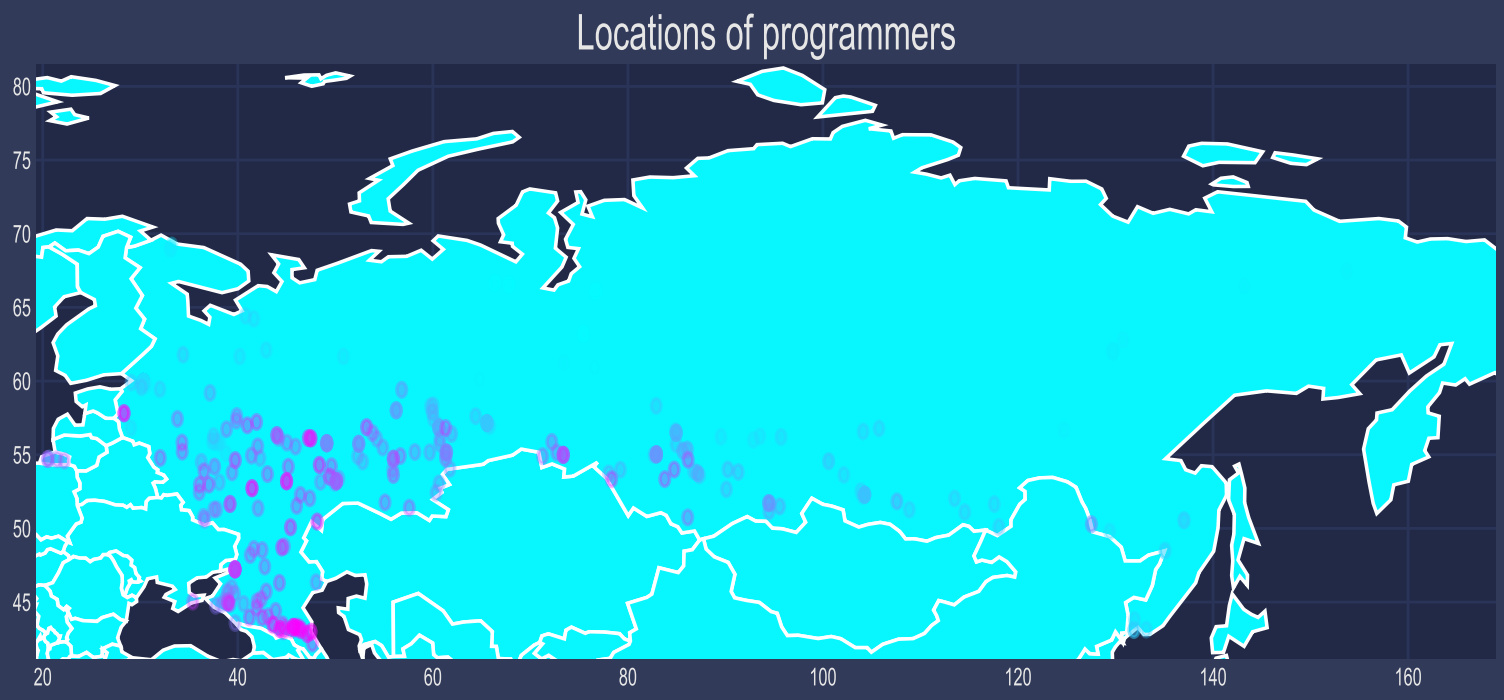
I’ve drastically modified 2 features. Language_knowledge became english_level and work_exerience became n_jobs. But that is not enough for the model, so we can create new features.

The feature that will not play any role in the model but would be heplful for EDA is coordinates of hometowns of programmers. I got them by using additional ‘КЛАДР’ dataset and web API.

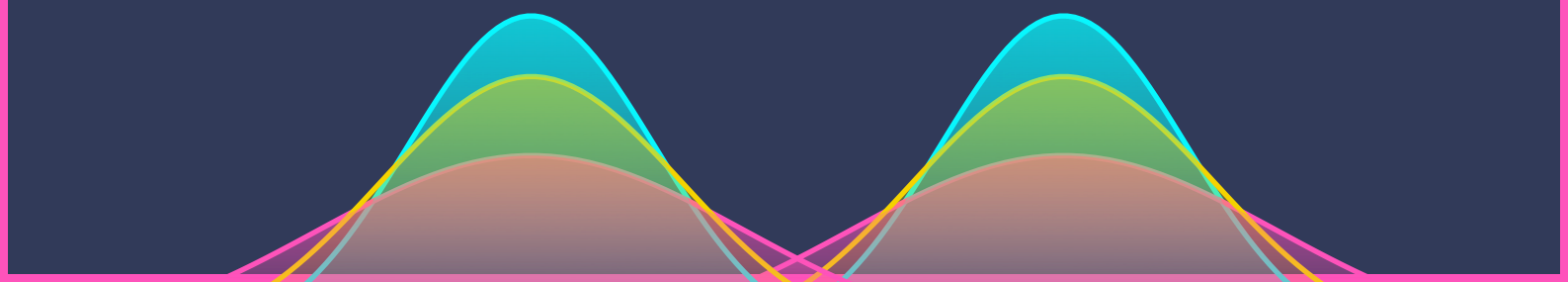
Also, I’ve created a helpful feature called mean_region_salary that contains information conforming to the name of the column. I’ve curated this dataset from officials, so there are some questions to it’s representativeness, but that’s the best we can have. The feature will conribute to the prediction with its big numbers.

EDA

Let’s see how the programmers are distributed across Russia

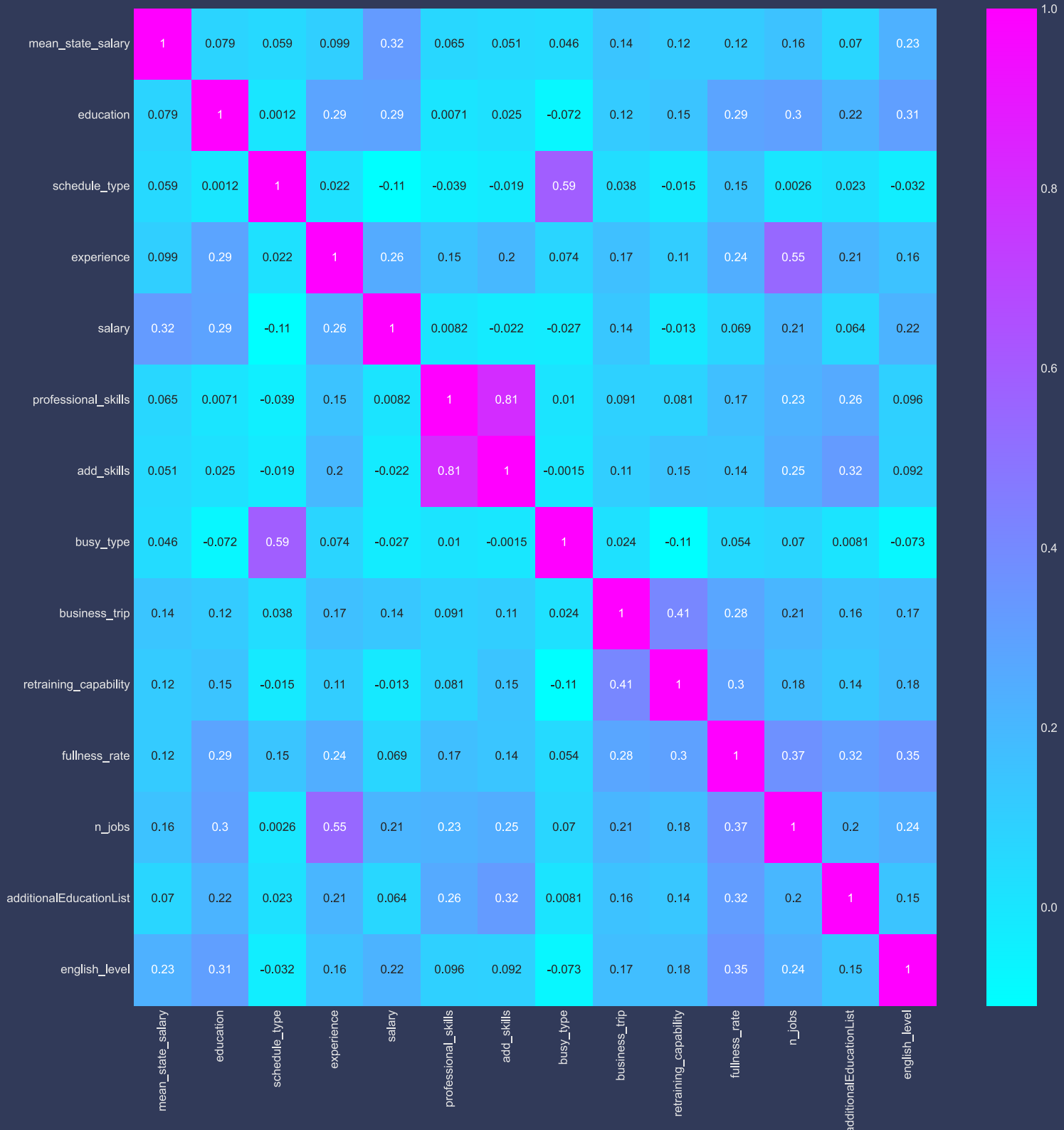


We can see that most submissions come from southern regions of Russia, which are not that rich. Therefore, low salaries for IT-professionals in our dataset are kinda justifiable



Heatmaps

The heatmap of all the columns in resulting dataset

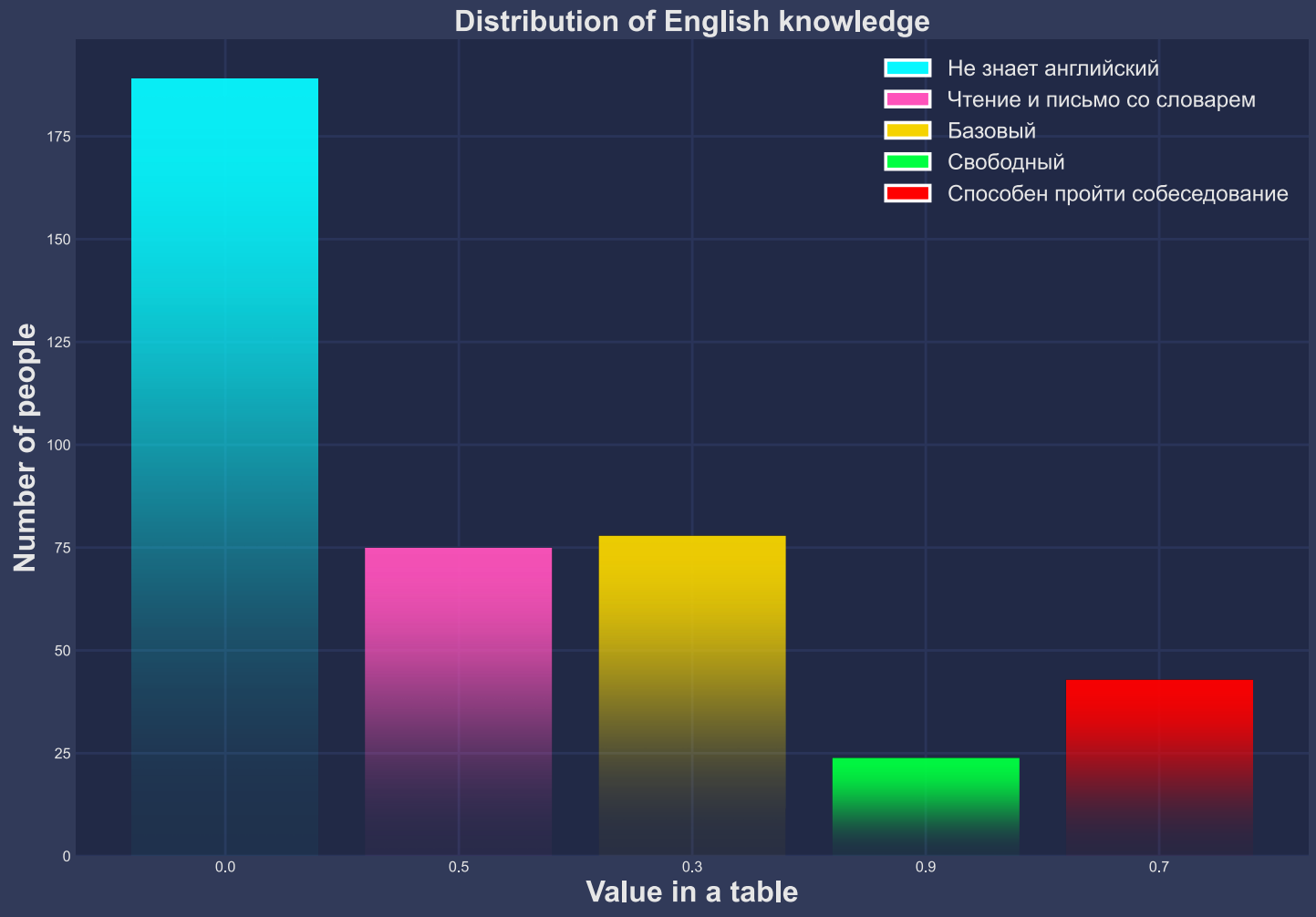


Heatmaps

If we throw away all the variables that almost do not correlate with the target we get this:

English

I consider English level to be one of the most vital for IT-professionals. So, it is good to look at barplots on this topic.



We can notice that a large group of people doesn’t know English. And only tiny amount of respondents are capable of acing the interview in English

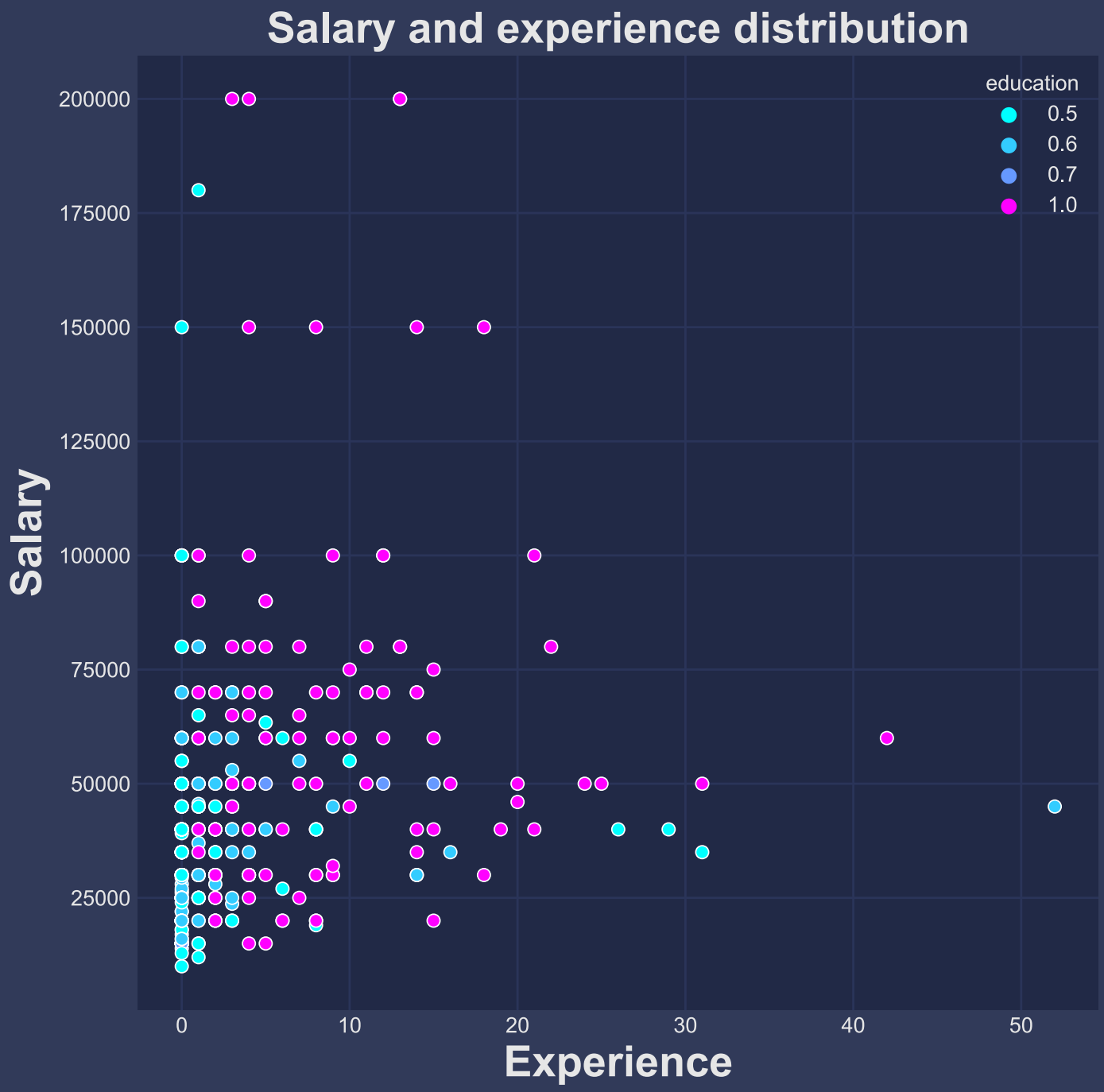
English

Since this skill is vital and not many people actually have it, it is easy to suppose that English level will correlate with the salary quite well.

There is definetely a trend: the better your English is, the better the salary gets. There are some fluctuations but I guess it is because of the dataset size.

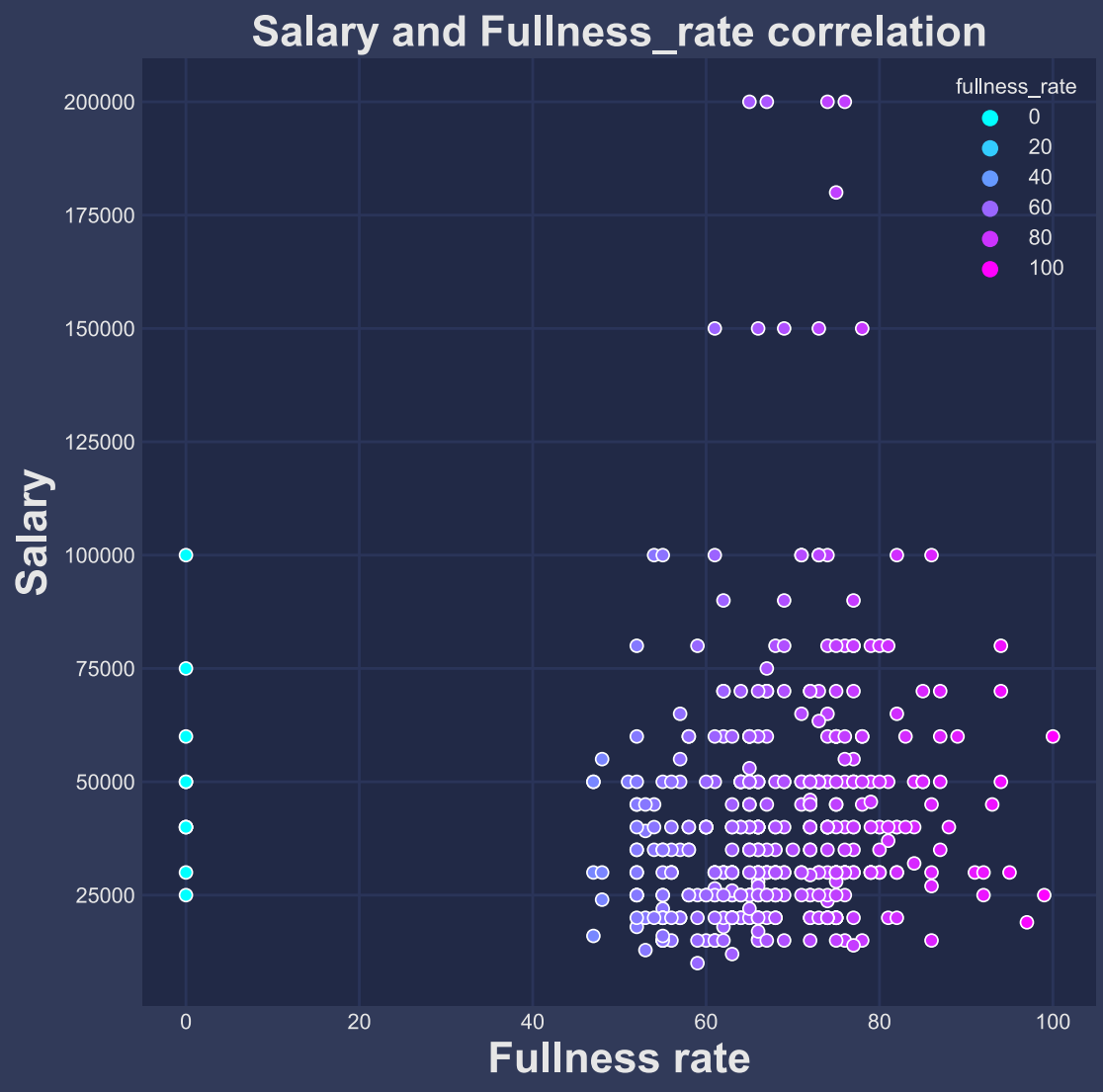
Education

From this plot we can see that better education mostly leads to more experience but doesn't neccessarly affect the salary



Fullness rate

That value shows how much information there is in your cv. It surely do positively affect salary, since there is not many big salaries without a decent cv on the plot



Model design

I’ve decided to use CatBoostRegressor for that problem. The main reason being its familiarity to me, but also I’ve considered that CatBoost is generally better performing on small datasets with fast learning than other algorithms like LGBM or XGB.

For the loss function I’ve chosen RMSE. There was an attempt to use MSLE, to compensate for anomalies, but I don’t like the tendency of this function to give more penalty for underscoring.

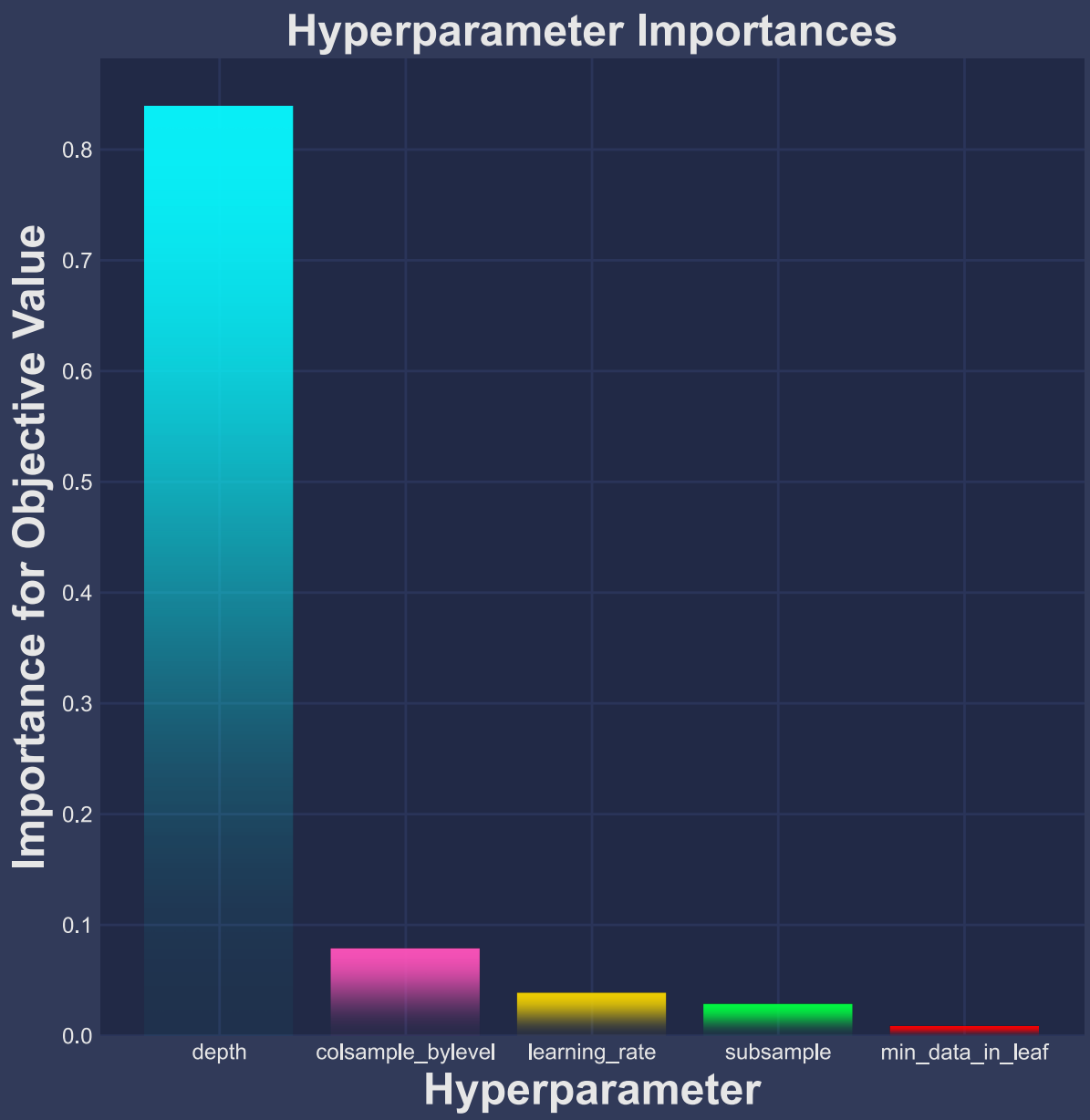
Three studies

I was finding the optimal hyperparamters for the model using Optuna. I've tried different combinations of everything, including number of columns in dataset. As it turns out, there is little difference between model that is fit on 15 columns, on 7 columns and on 4 columns.

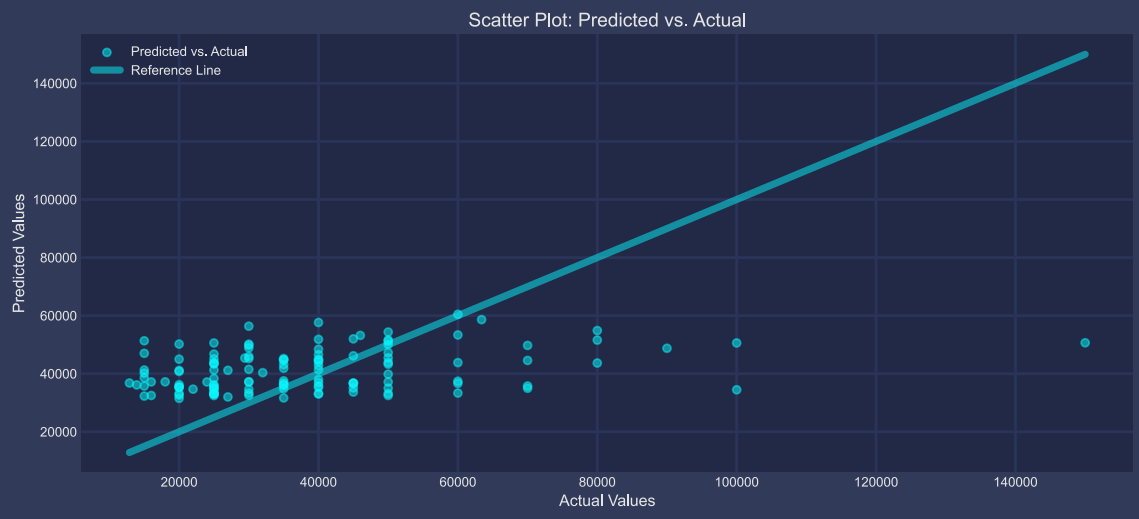
But for every number of columns I've left a separate model and study in notebook. Maybe we will see a big difference in them when it'll be introduced to a big dataset.

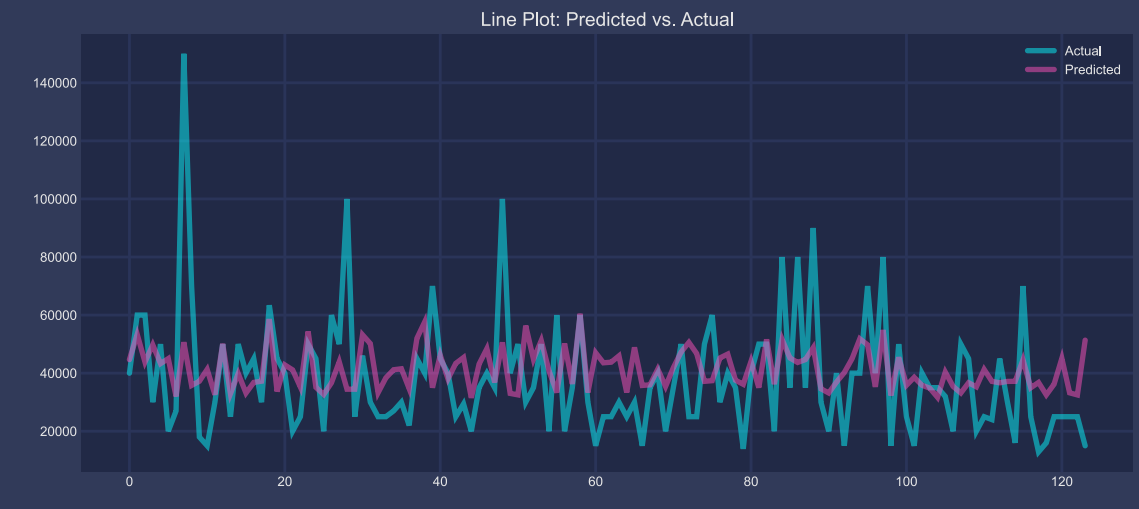
Hyper Parameters

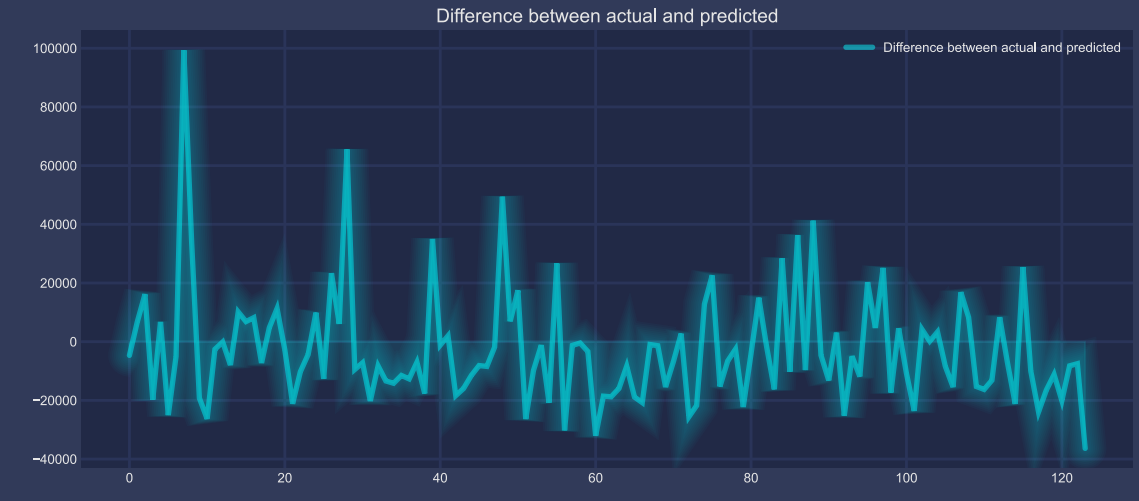
During those studies I've tracked the importances of each hyperparameter. The picture is more or less the same everywhere



15-column model

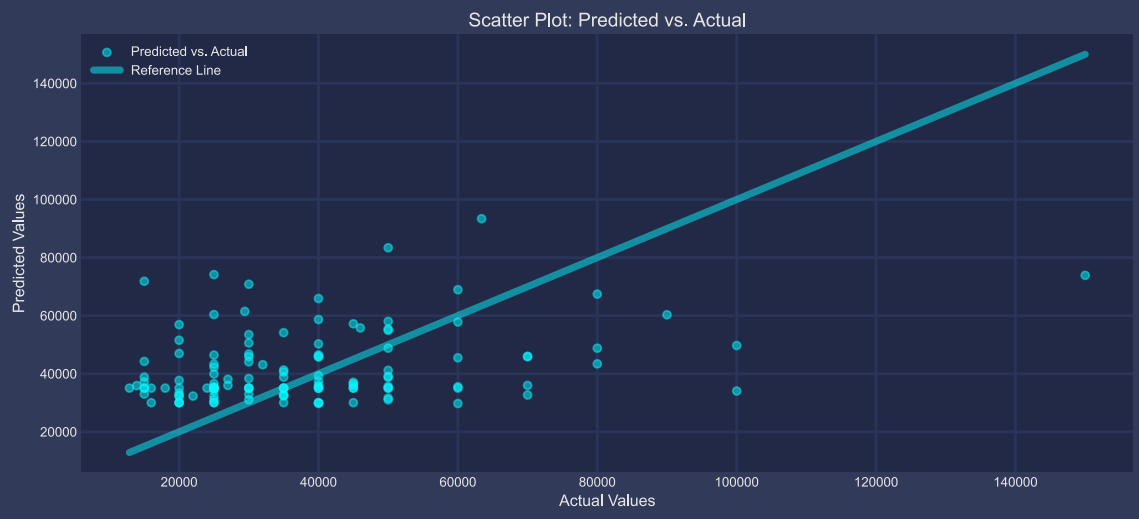


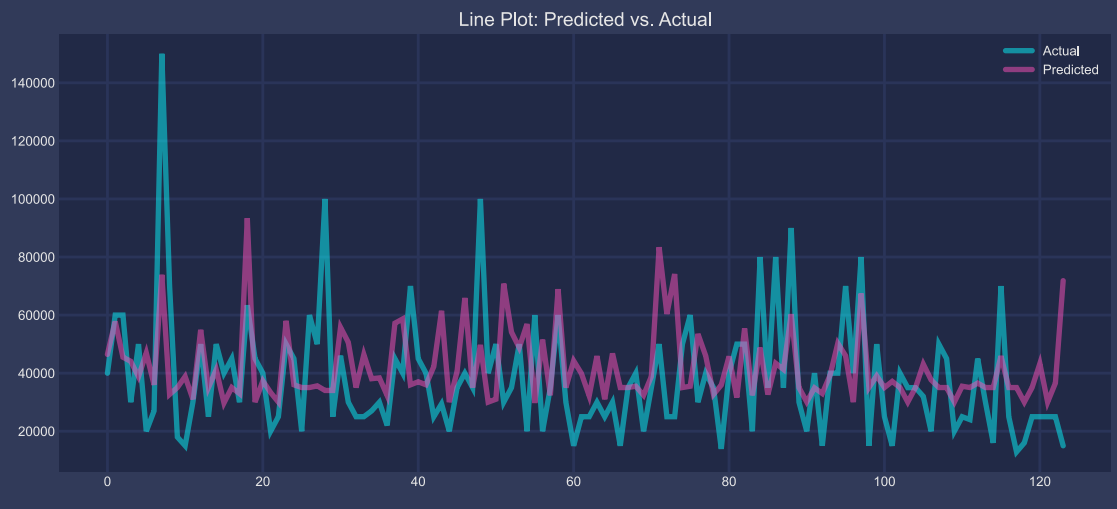


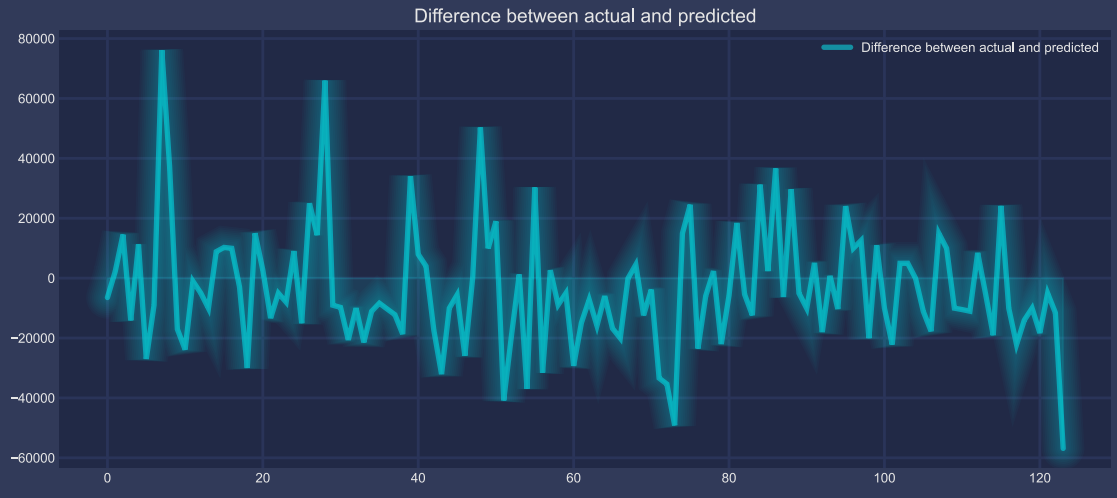


RMSE: 19656

7-column model

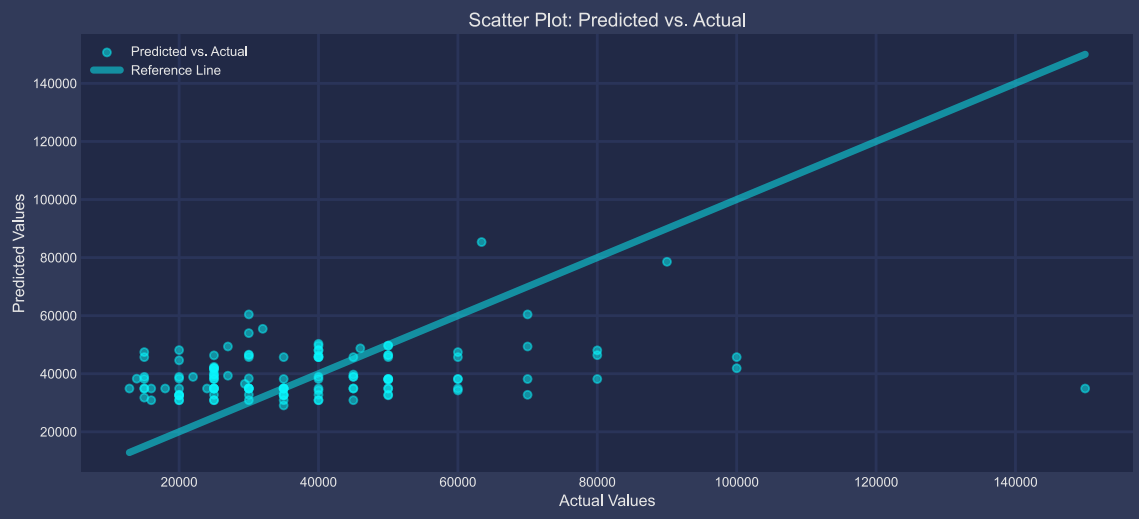


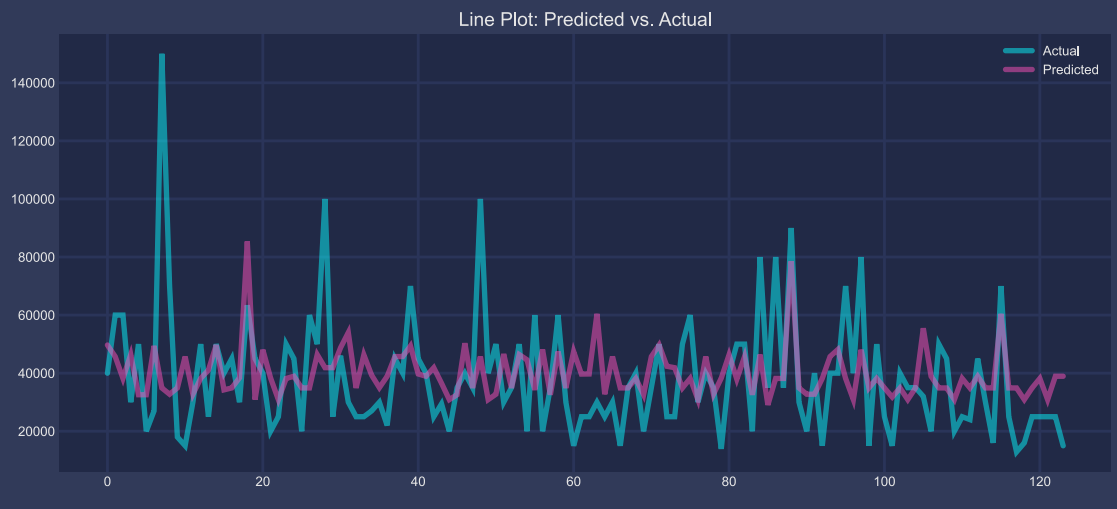


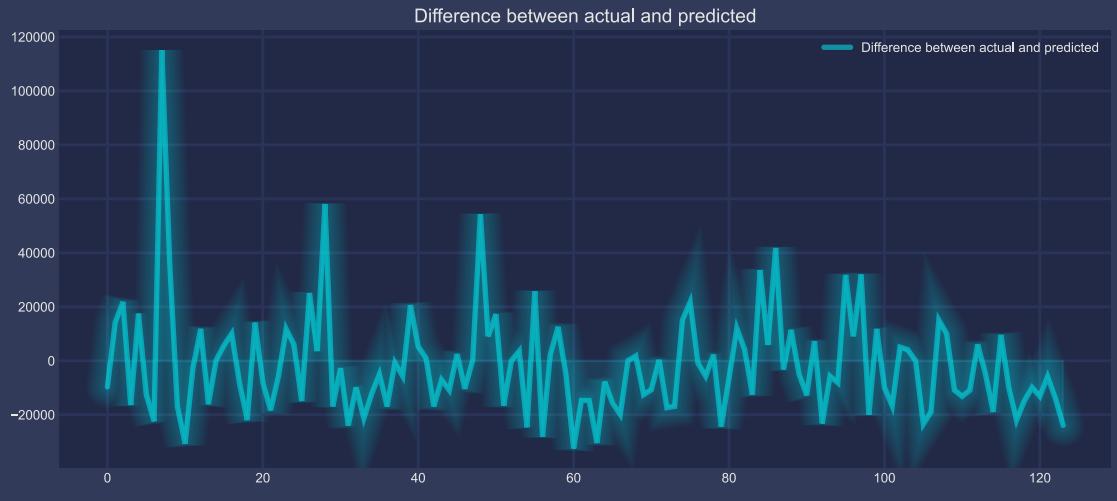


RMSE: 20585

4-column model







RMSE: 19986

Conclusion

I think I've done a decent job analyzing this dataset and building the model. The only thing it lacks is the size. It had a lot of coluns, but clearly not enough rows to train and evaluate the results of the model.

By the way, if you input data about me in the models, they throw the value around 60k. I think it is overestimating, as we can see from high RMSE, but the value is looks great.

Thank you for reading through the record and don't forget to checkout the notebook afterwards.

Goryachev Alexander
27.07-01.08