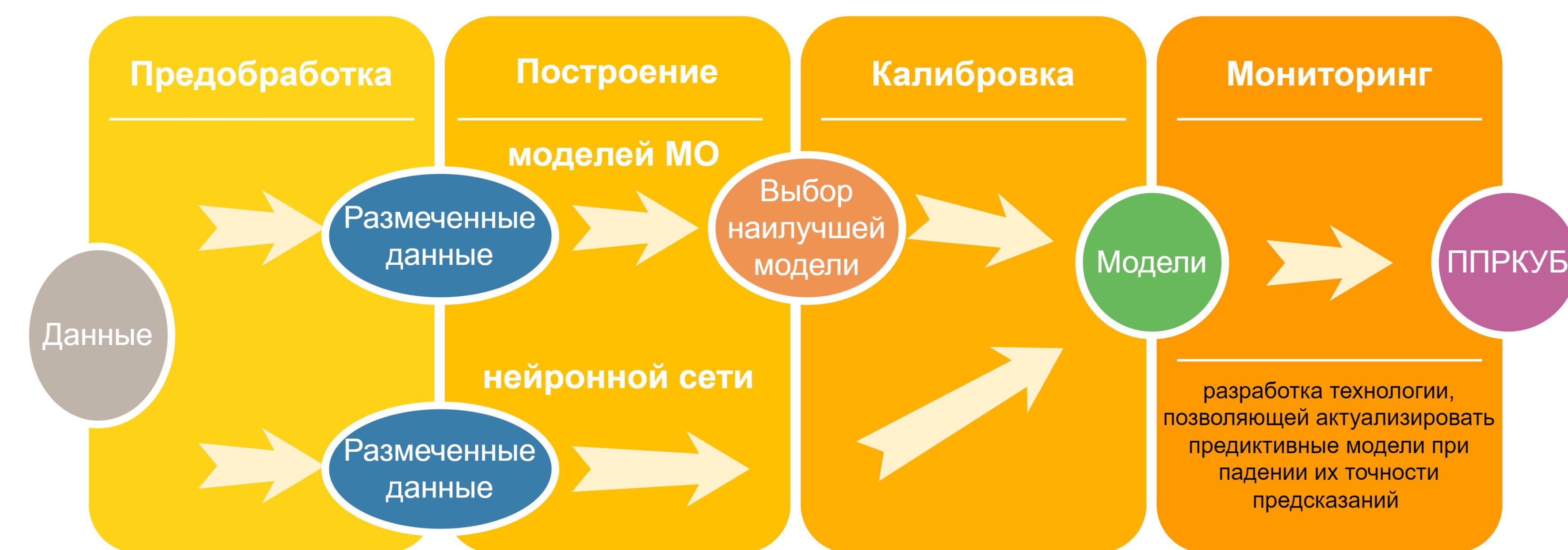


ЭТАПЫ РАЗРАБОТКИ ПРОГРАММНОЙ ПОДСИСТЕМЫ РАСЧЕТА КЛИЕНТСКОЙ УБЫЛИ БАНКА

Предварительная обработка данных

- Трансформация категориальных переменных
- Отсечение линейнозависимых признаков
- Обработка пропущенных значений
- Нормализация данных
- Удаление выбросов
- Разбиение выборки на обучающую, тестовую и валидационную



Выбор модели машинного обучения

- Классические модели машинного обучения
- Ансамбли моделей
- Бэггинг, Стэкинг и Бустинг
- Библиотеки градиентного бустинга

Применение нейронных сетей

- Общие архитектуры нейронных сетей
- Построение нейронной сети

Калибровка полученных моделей

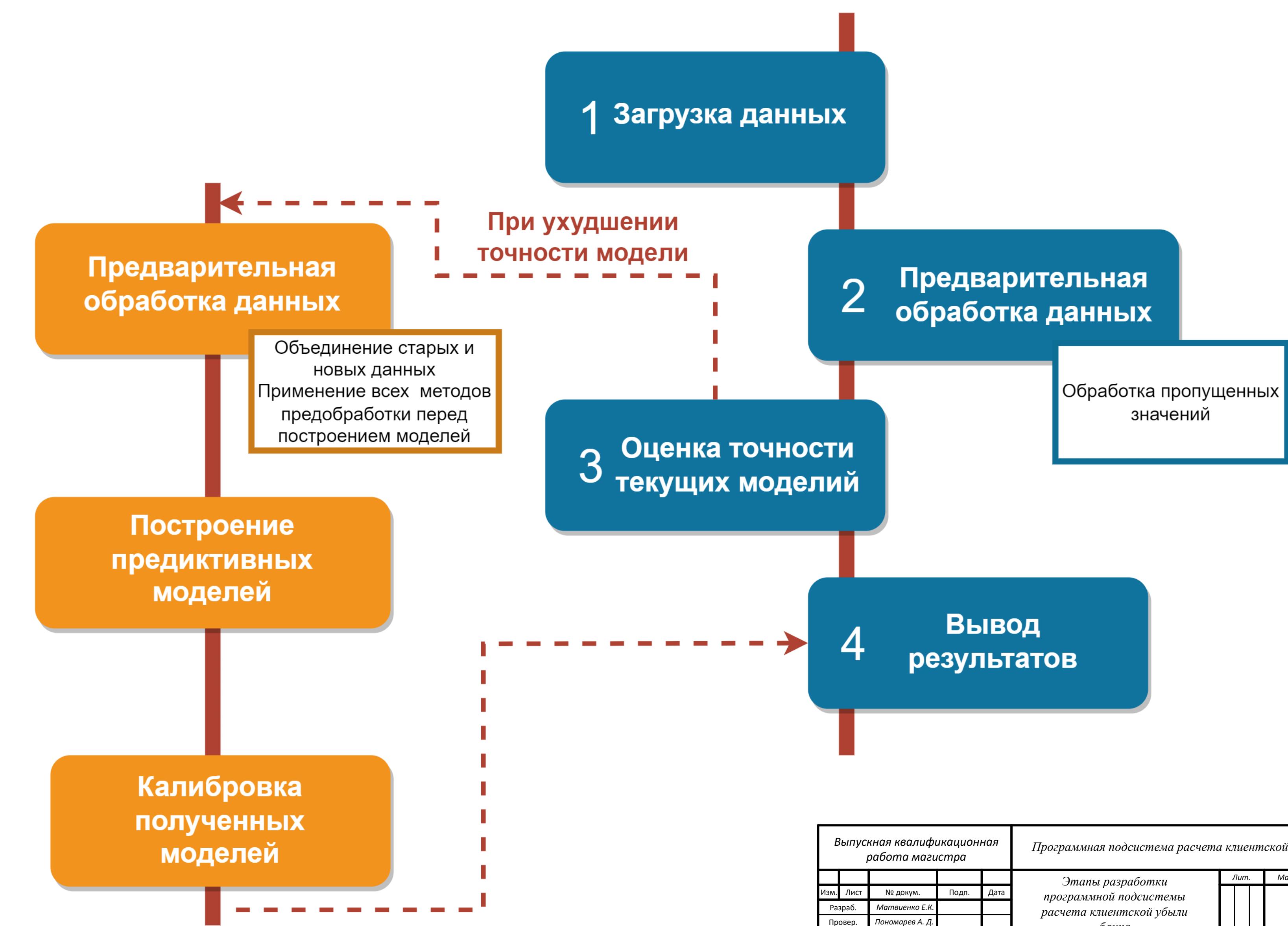
- Построение кривых надежности
- Расчет метрики Brier Score для моделей машинного обучения

Вывод результатов работы моделей

- Таблицы с указанием вероятности ухода клиентов
- Построение AUC-кривых
- Графики, указывающие на наиболее значимые параметры и их интерпретация

Разработка технологии мониторинга

- Алгоритм технологии мониторинга
- Построение графика мониторинга работоспособности моделей

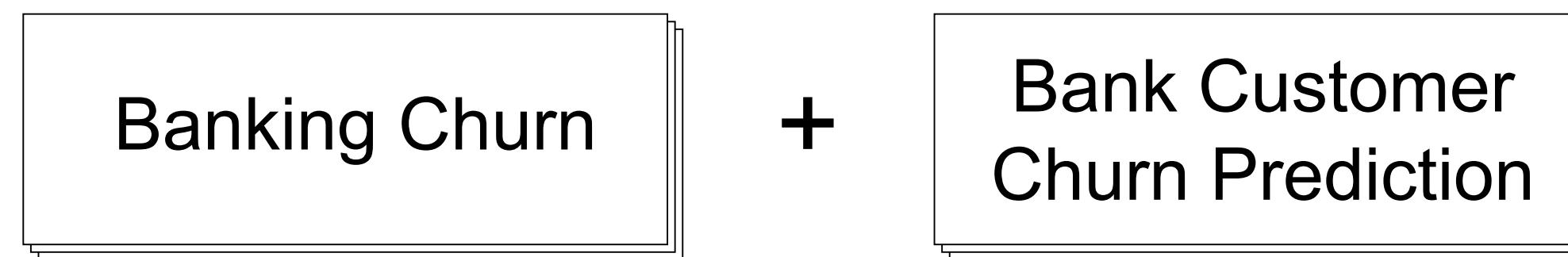


Выпускная квалификационная работа магистра					Программная подсистема расчета клиентской убыли банка			
Изм.	Лист	№ докум.	Подл.	Дата	Этапы разработки программной подсистемы расчета клиентской убыли банка	Лит.	Масса	Масштаб
Разраб.	Матвеенко Е.К.							
Правер.	Пономарев А. Д.							
Н.контроль	Миннитаева А.М.							

MГТУ им. Н.Э. Баумана
Кафедра ИУБ
Группа ИУ6-43М

ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ

Объединение датасетов



Удаление выбросов

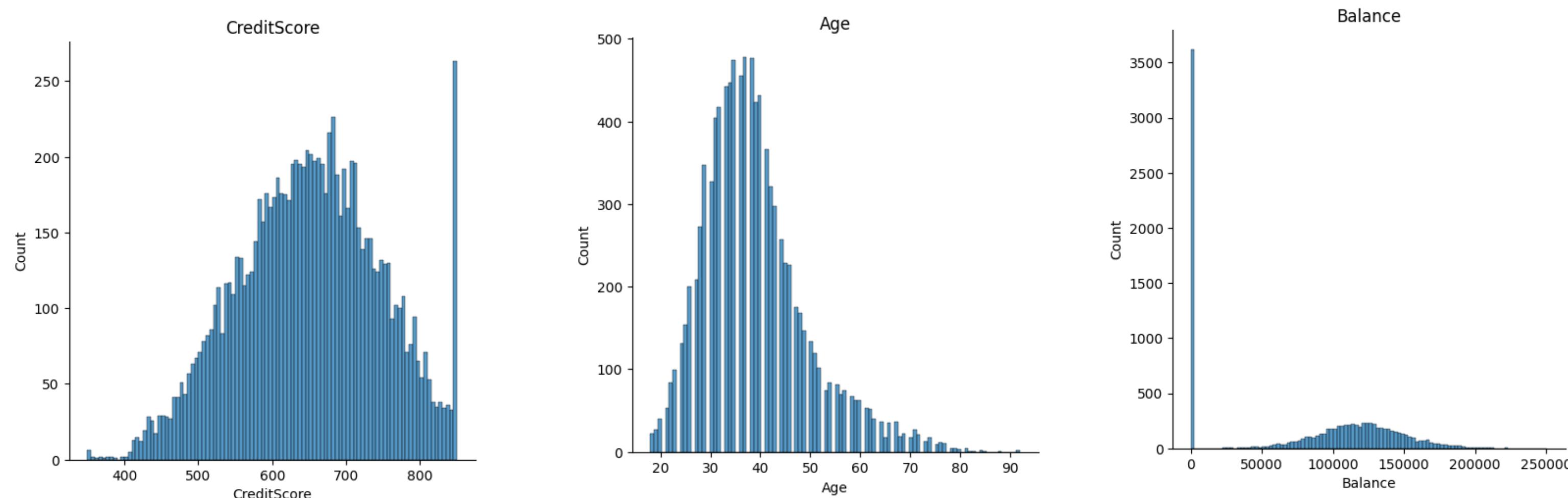
$$\text{IQR} = Q_3 - Q_1$$

$$[Q_1 - 1.5\text{IQR}, Q_3 + 1.5\text{IQR}]$$

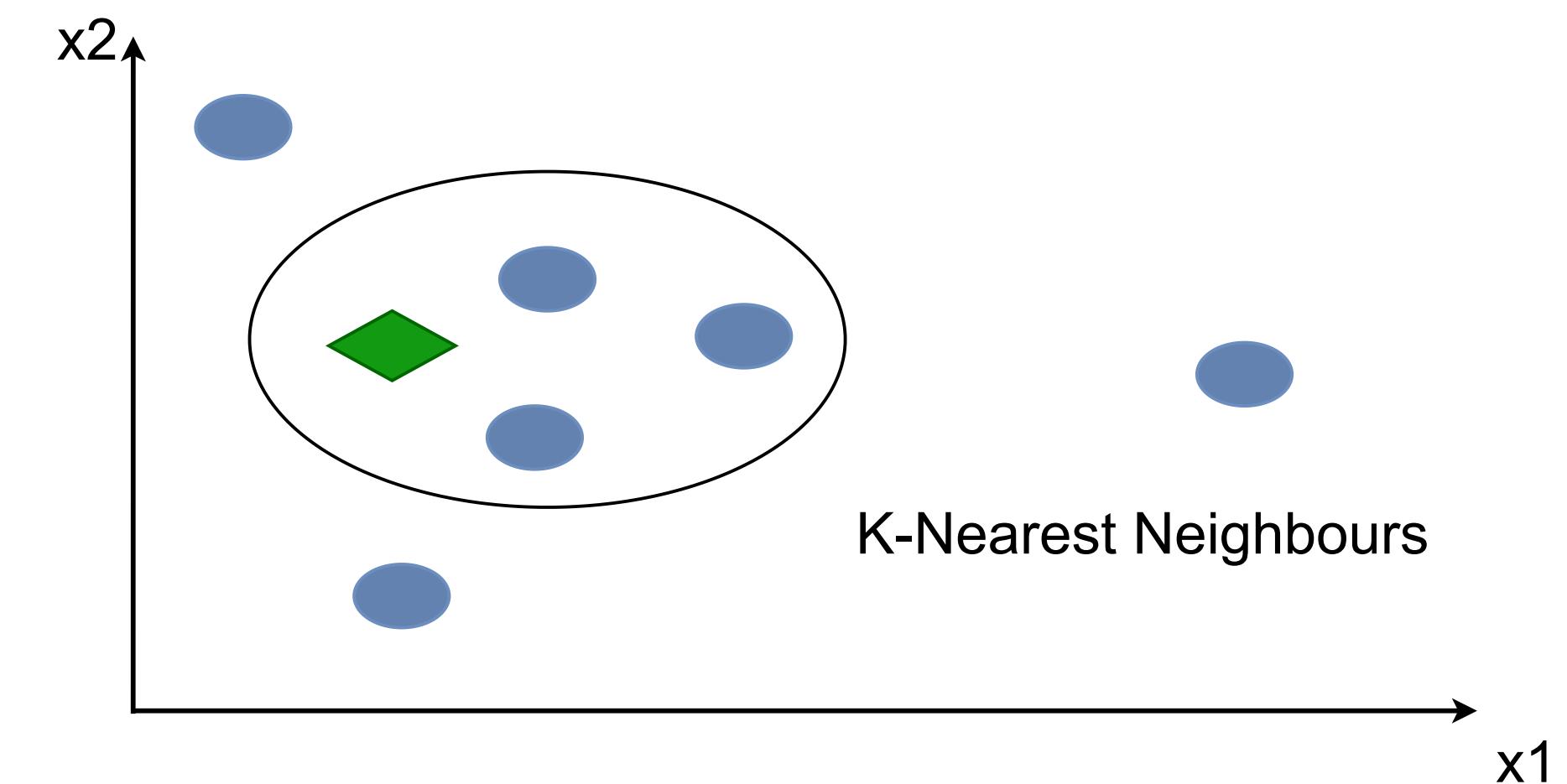
Нормализация данных

$$X_{\text{new}} = \frac{X_{\text{old}} - X_{\text{mean}}}{\sigma}$$

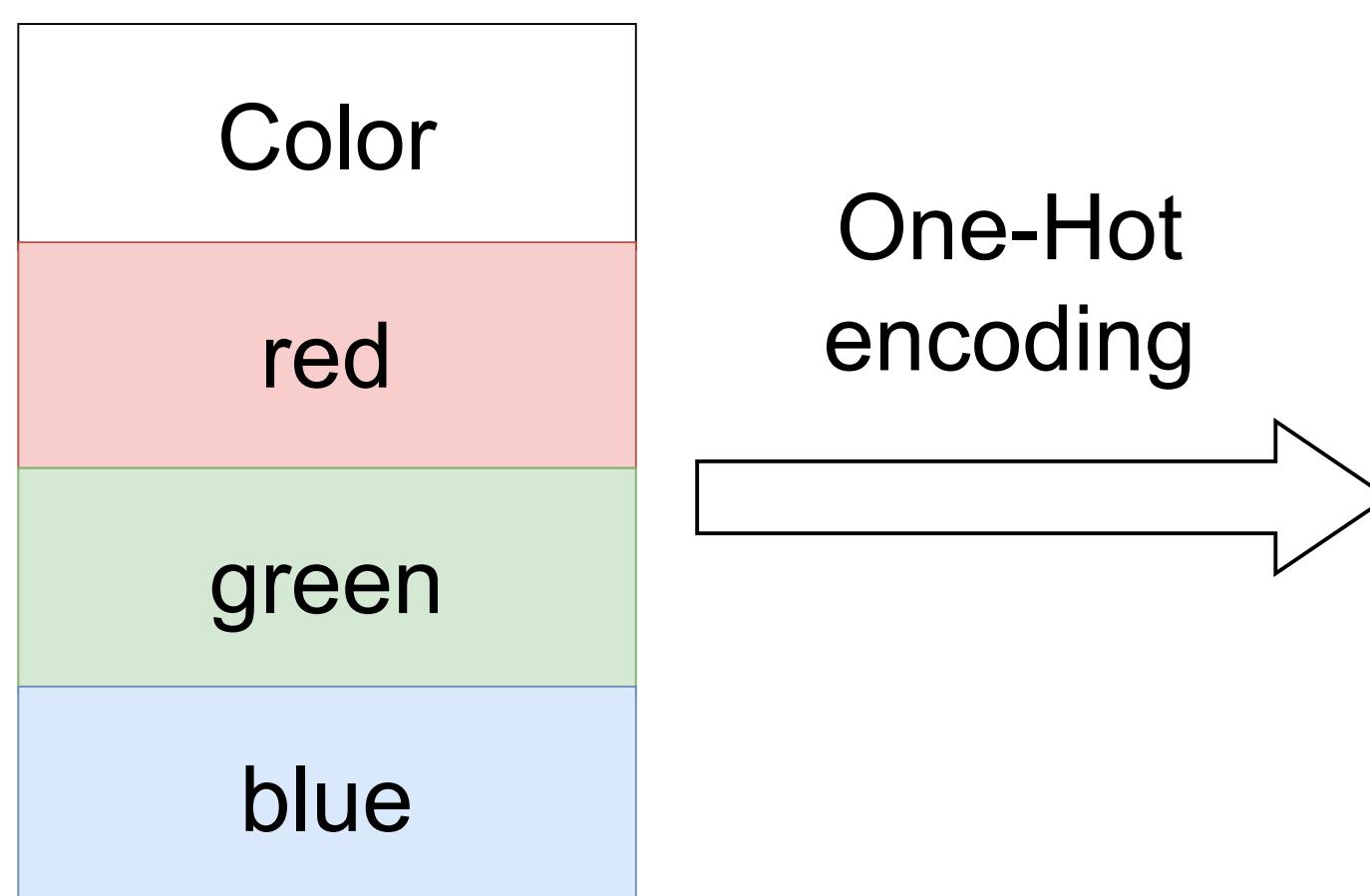
Визуальный осмотр данных



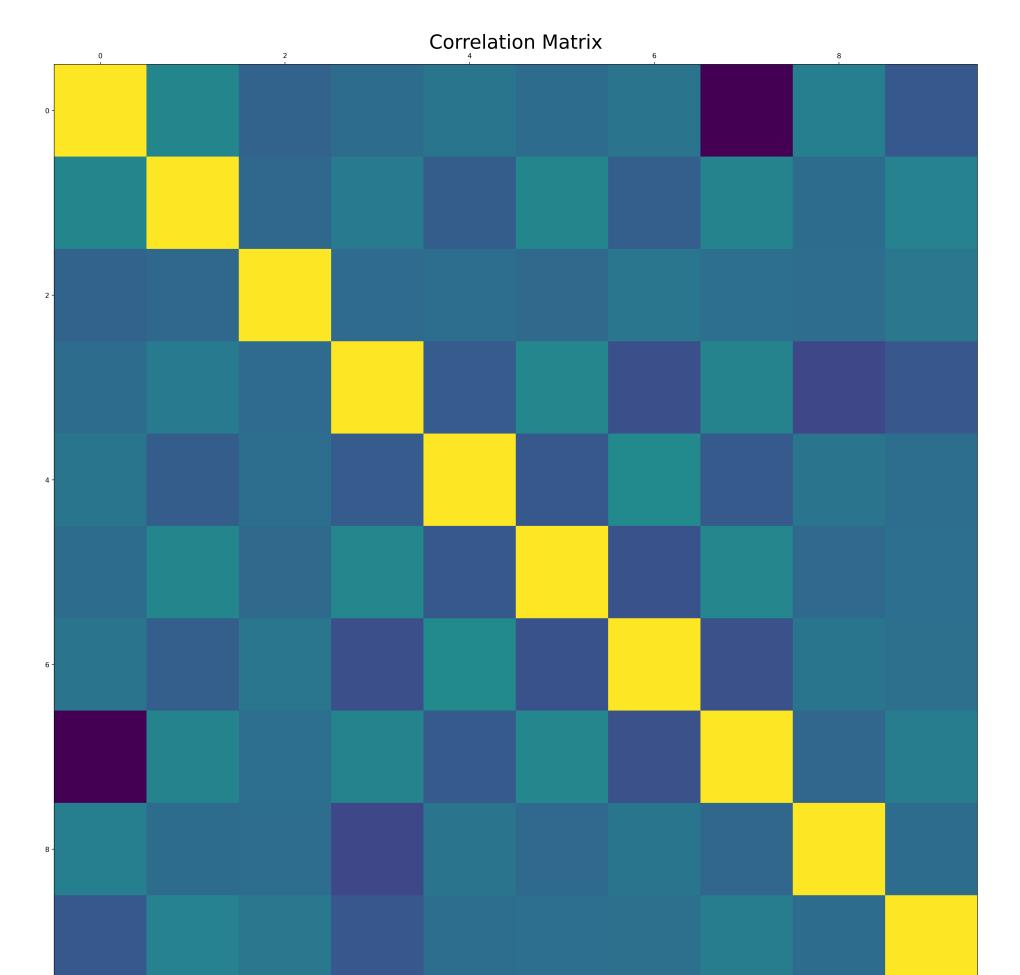
Обработка пропущенных значений



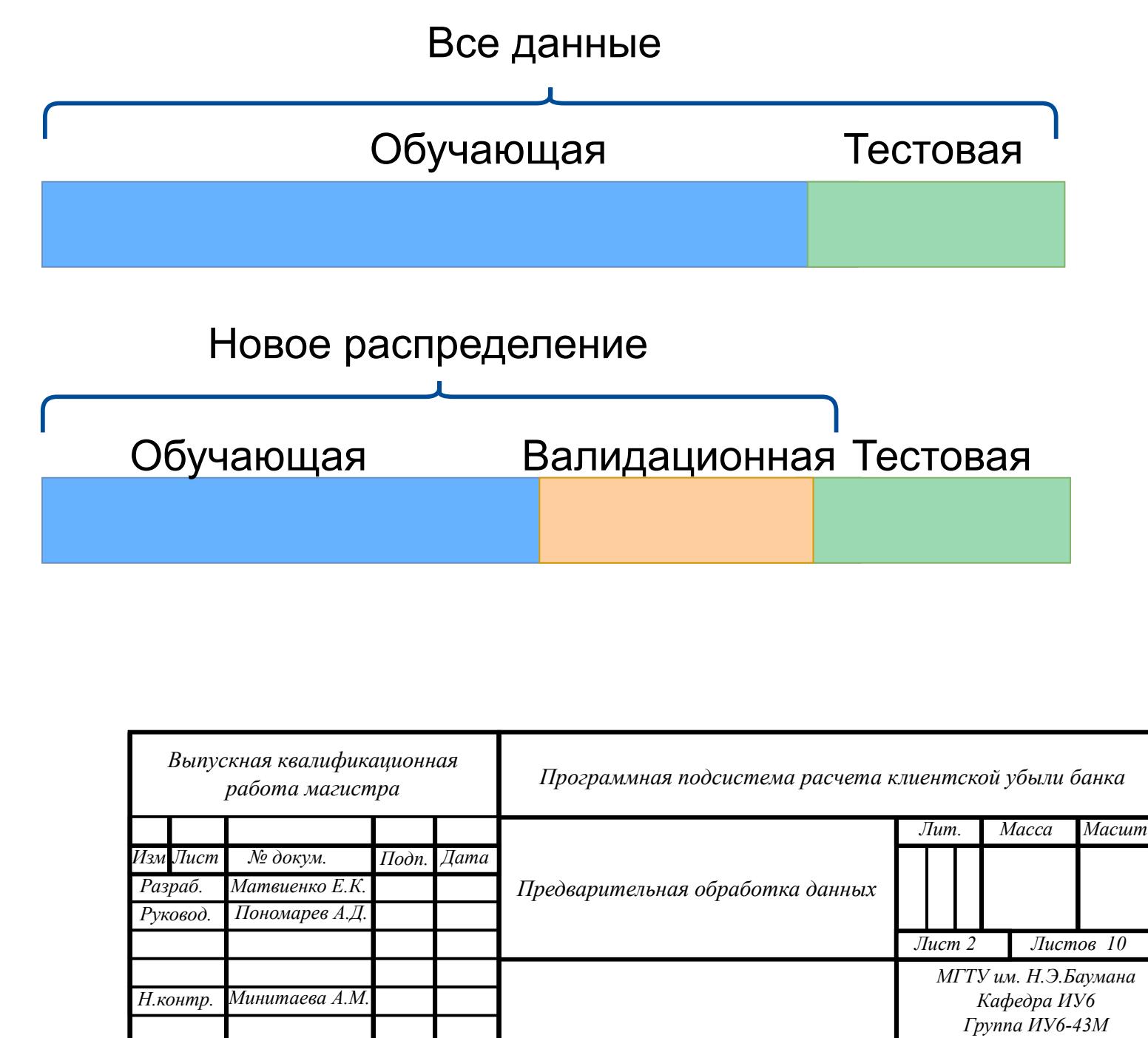
Трансформация категориальных переменных



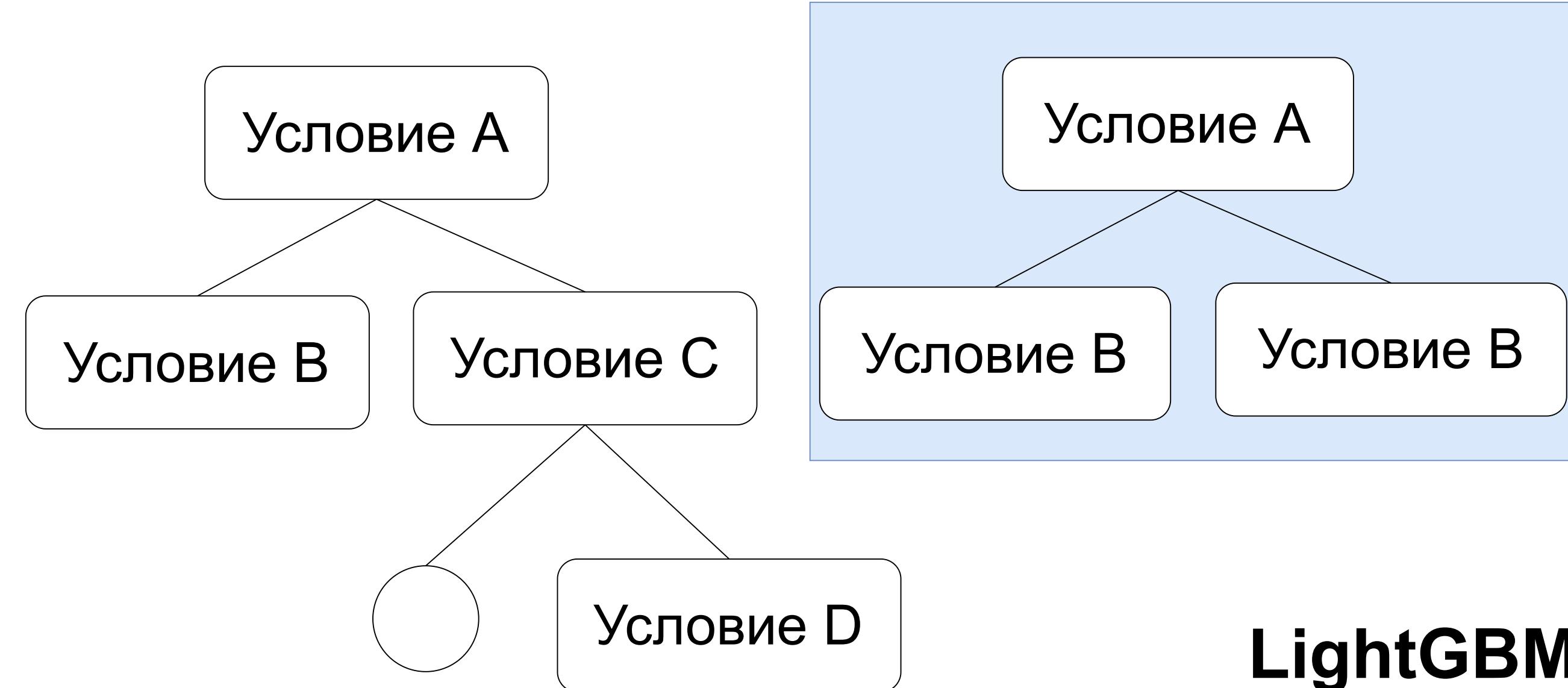
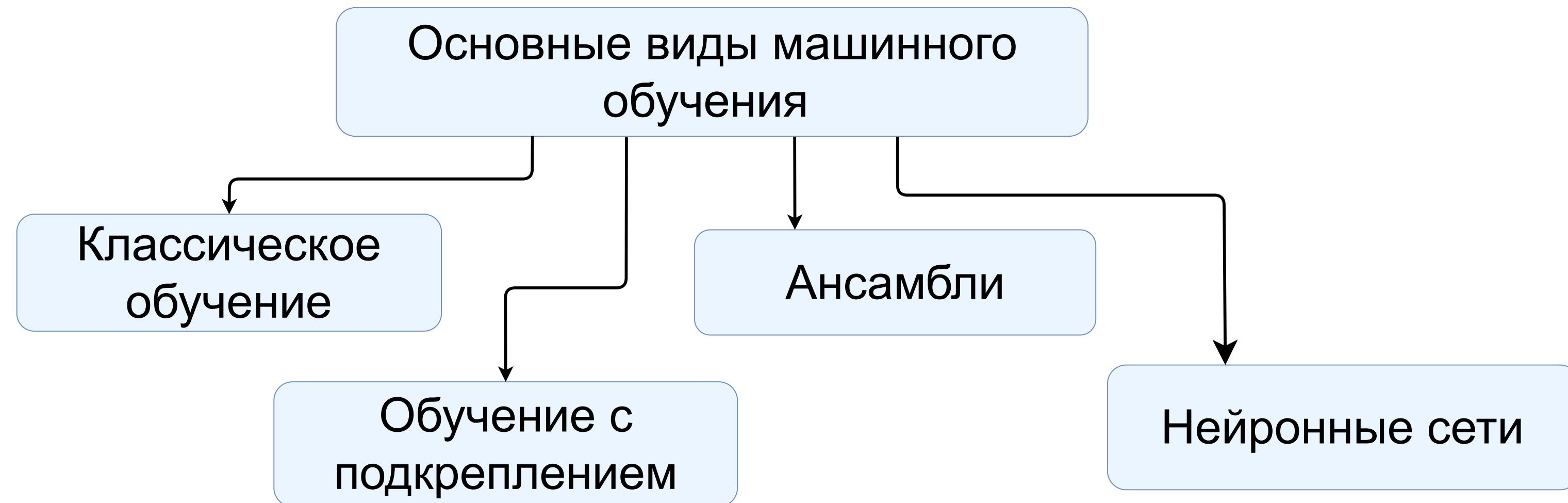
Отсечение линейнозависимых признаков



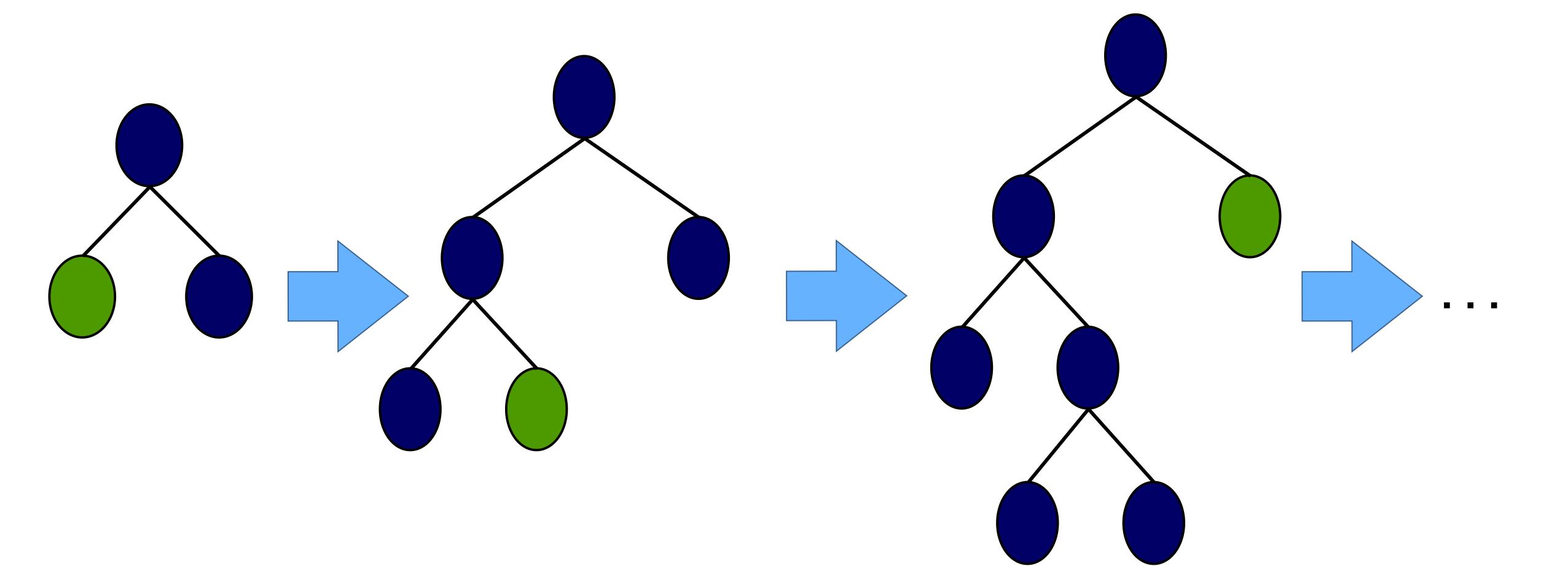
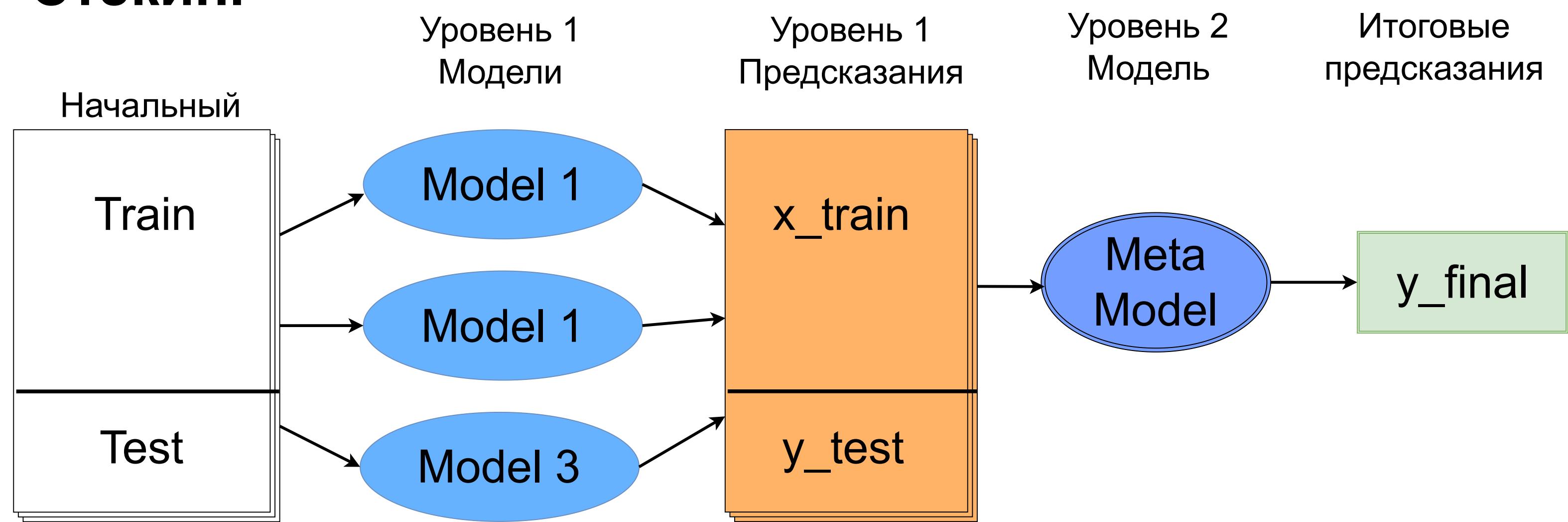
Разбиение выборки



АНАЛИЗ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

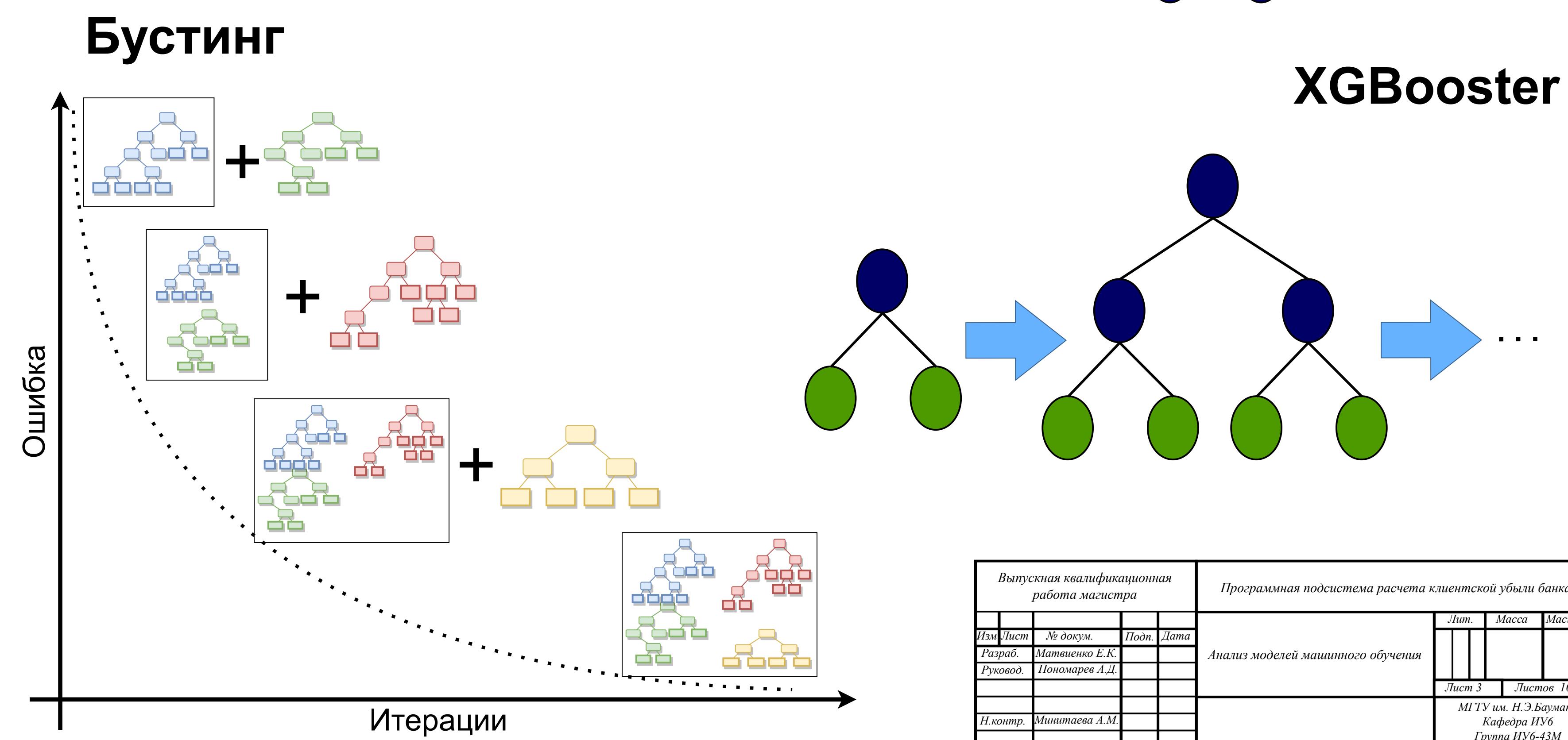
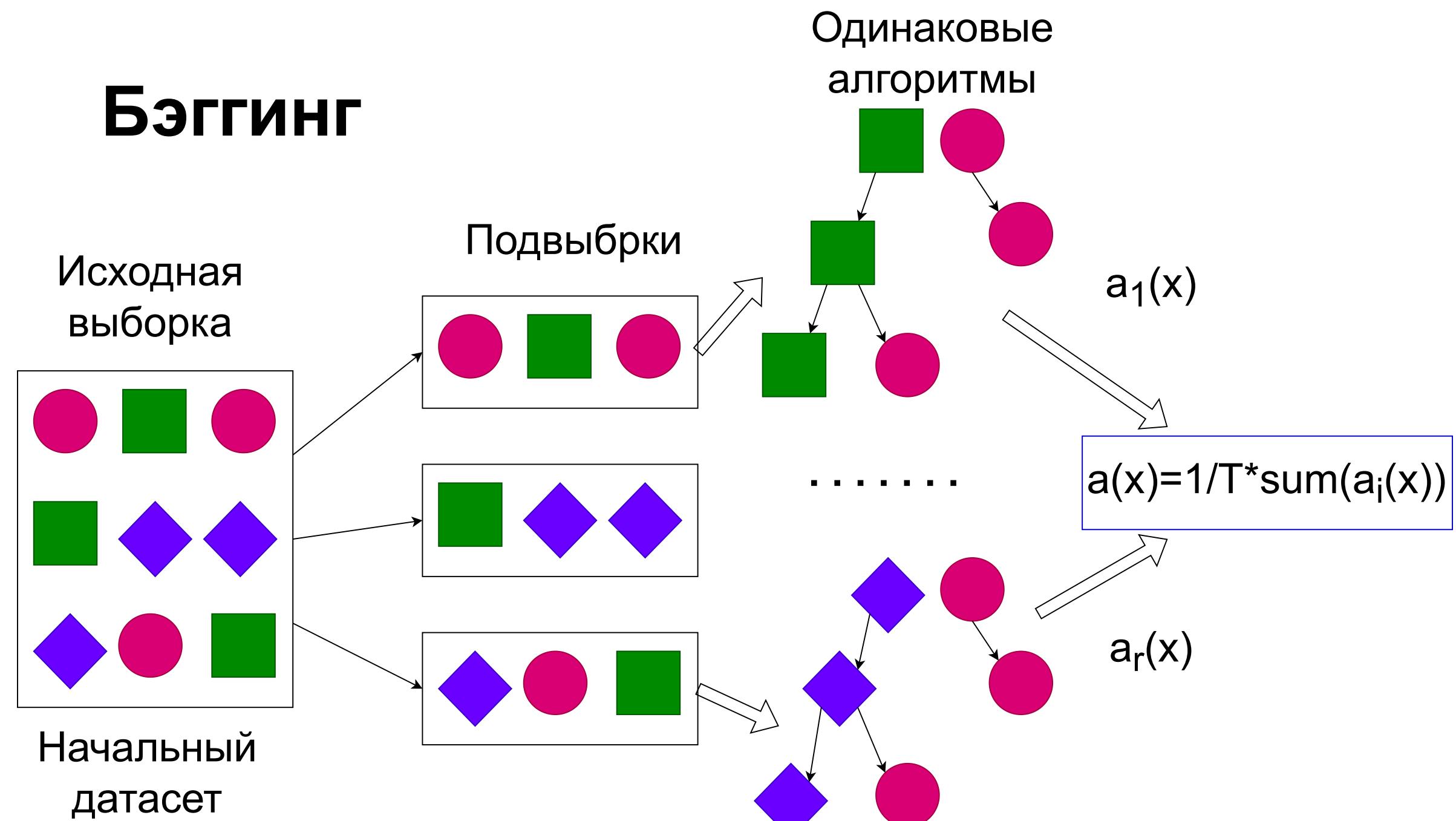


Стэкинг



CatBoost

БЭГГИНГ



ПОСТРОЕНИЕ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ И ВЫБОР НАИБОЛЕЕ ТОЧНОЙ МОДЕЛИ

XGB booster

```
params_xgb = {
    'n_estimators': np.arange(1, 100, 25),
    'max_depth': np.arange(1, 10, 3),
    'max_leaves': np.arange(0, 20, 3),
    'grow_policy': [0, 1],
    'learning_rate': np.arange(0.1, 1, 0.3),
    'booster': ['gbtree', 'gblinear', 'dart'],
    'subsample': np.arange(0.6, 0.8, 0.1),
    'colsample_bytree': np.arange(0.6, 0.8, 0.1),
    'colsample_bylevel': np.arange(0.6, 0.8, 0.1),
    'colsample_bynode': np.arange(0.6, 0.8, 0.1),
    'reg_alpha': np.arange(0, 1, 0.5),
    'reg_lambda': np.arange(0, 1, 0.5),
}
```

LGBM booster

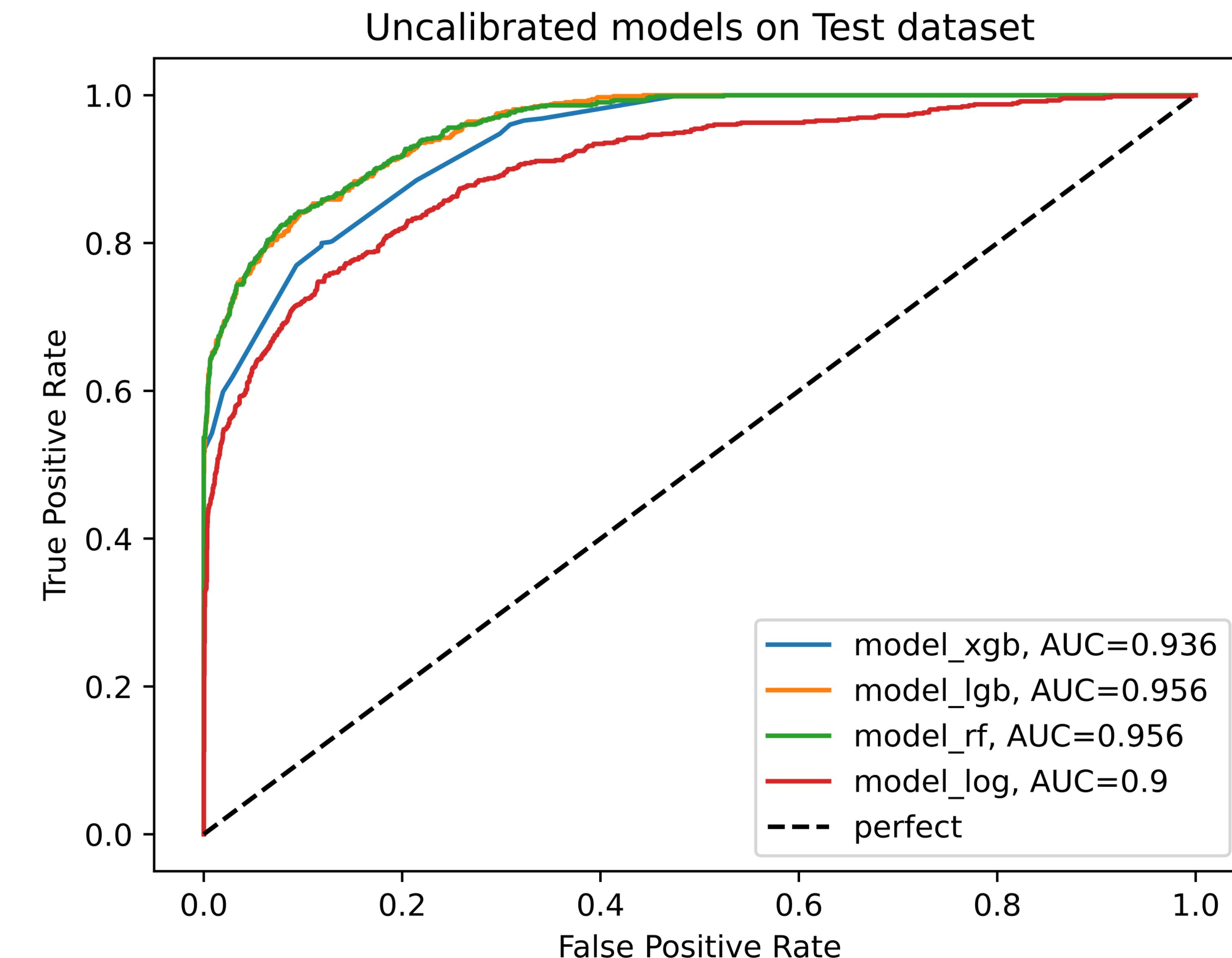
```
params_lgb = {
    'n_estimators': np.arange(1, 100, 25),
    'max_depth': np.arange(1, 10, 3),
    'num_leaves': np.arange(0, 20, 3),
    'learning_rate': np.arange(0.1, 1, 0.3),
    'boosting_type': ['gbdt', 'goss', 'dart'],
    'subsample': np.arange(0.6, 0.8, 0.1),
    'colsample_bytree': np.arange(0.6, 0.8, 0.1),
    'reg_alpha': np.arange(0, 2, 0.5),
    'reg_lambda': np.arange(0, 2, 0.5),
}
```

XGB forest

```
params_rf = {
    'n_estimators': np.arange(1, 100, 50),
    'max_leaves': np.arange(0, 20, 10),
    'learning_rate': np.arange(0.1, 1, 0.3),
    'reg_alpha': np.arange(0, 1, 0.5),
    'reg_lambda': np.arange(0, 1, 0.5),
}
```

LogRegression

```
params_log = {
    'l1_ratio': np.arange(0, 1, 0.2)
}
```



- xgb booster precision metric: 0.76
- lgb booster precision metric: 0.78
- xgb forest precision metric: 0.82
- log regressor precision metric: 0.59

- xgb booster recall metric: 0.39
- lgb booster recall metric: 0.47
- xgb forest recall metric: 0.40
- log regressor recall metric: 0.17

- xgb booster f1 metric: 0.51
- lgb booster f1 metric: 0.59
- xgb forest f1 metric: 0.54
- log regressor f1 metric: 0.26

Программная подсистема расчета клиентской убыли банка					
Изм.	Лист	№ докум.	Подп.	Дата	
Разраб.	Матвиенко Е.К.				
Провер.	Пономарев А.Д.				
Н.контроль	Минихаева А.М.				

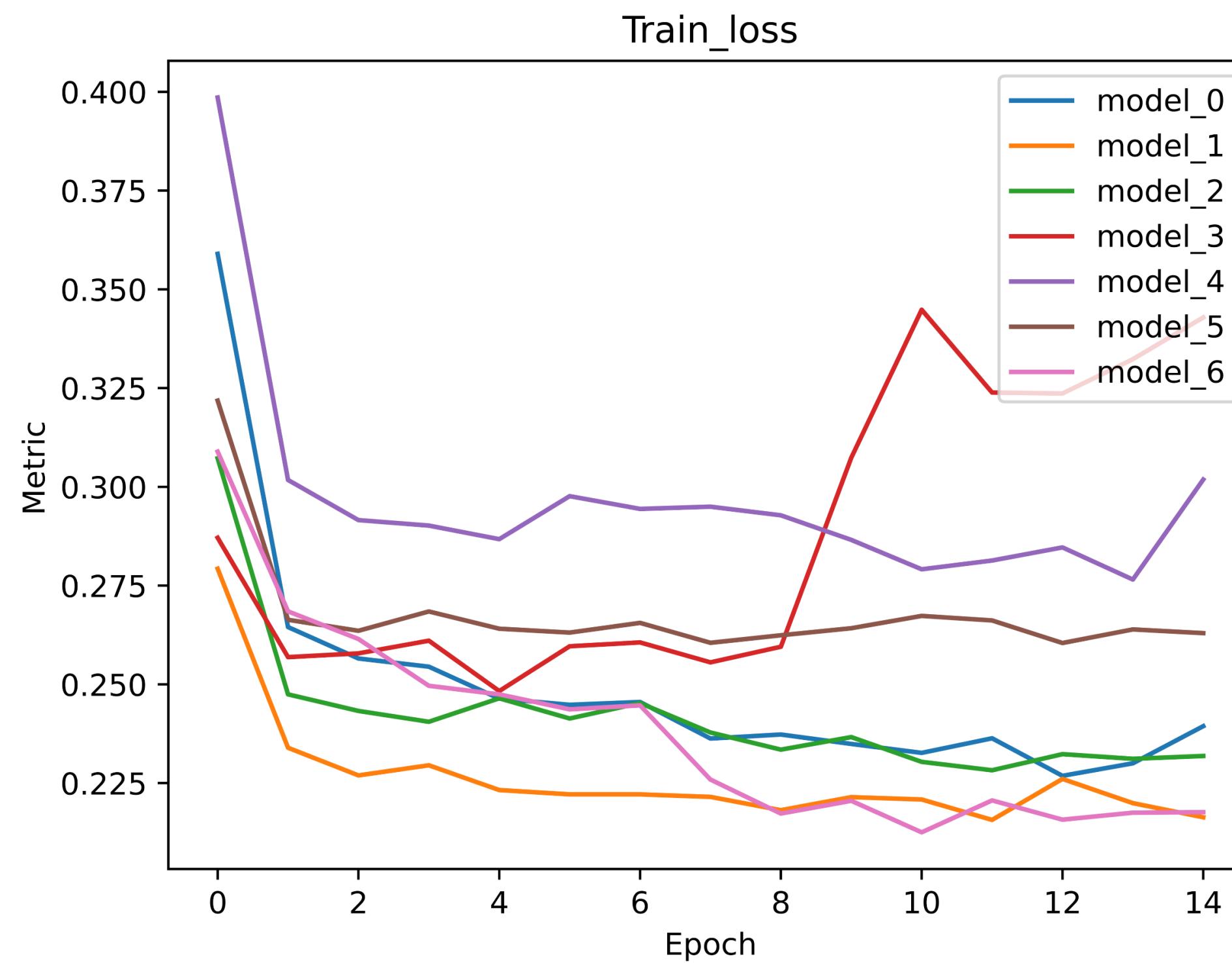
Построение моделей машинного обучения и выбор наиболее точной модели

Лист 4	Листов 10
--------	-----------

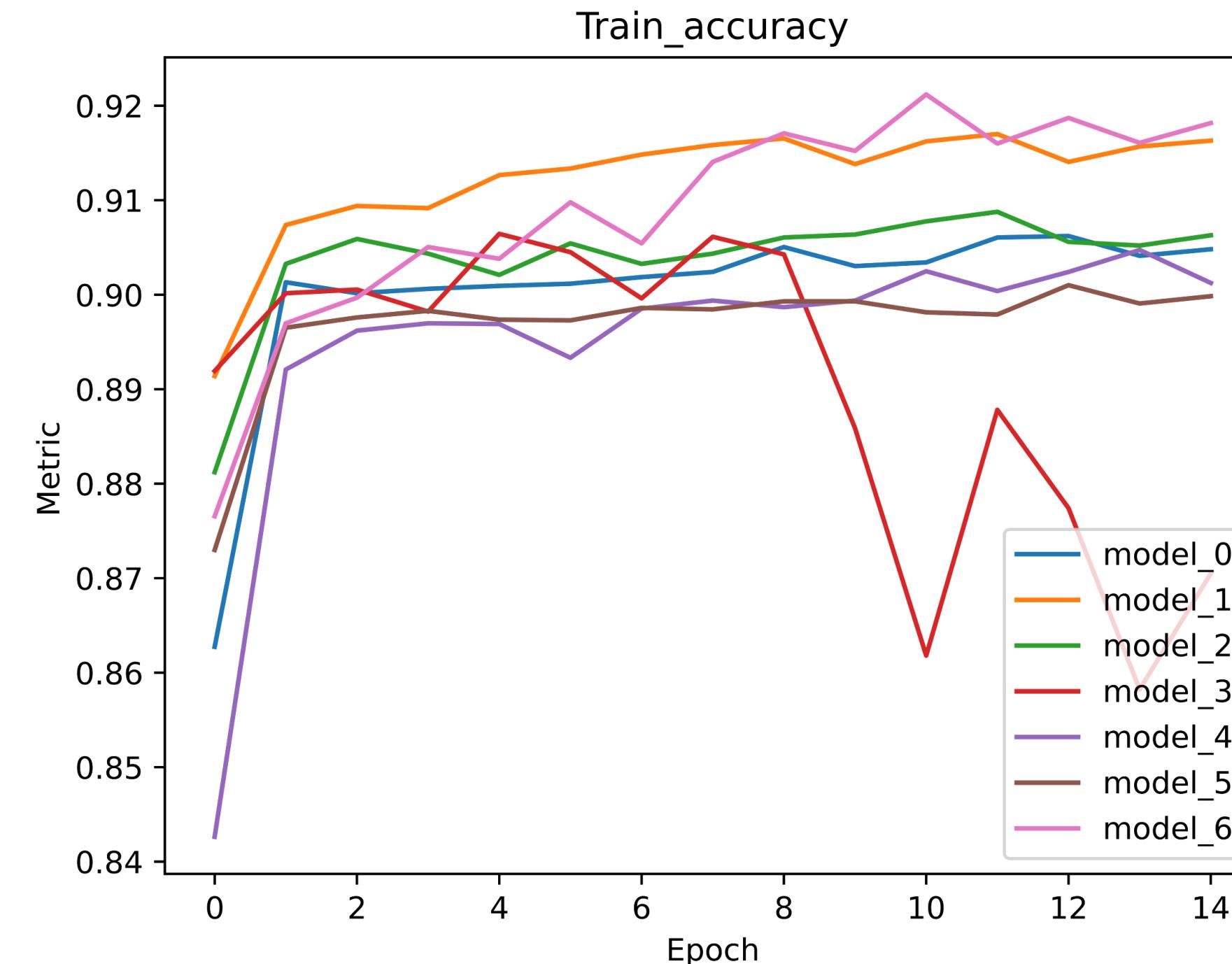
МГТУ им. Н.Э. Баумана
Кафедра ИУ6
Группа ИУ6-43М

ПОСТРОЕНИЕ НЕЙРОННЫХ СЕТЕЙ

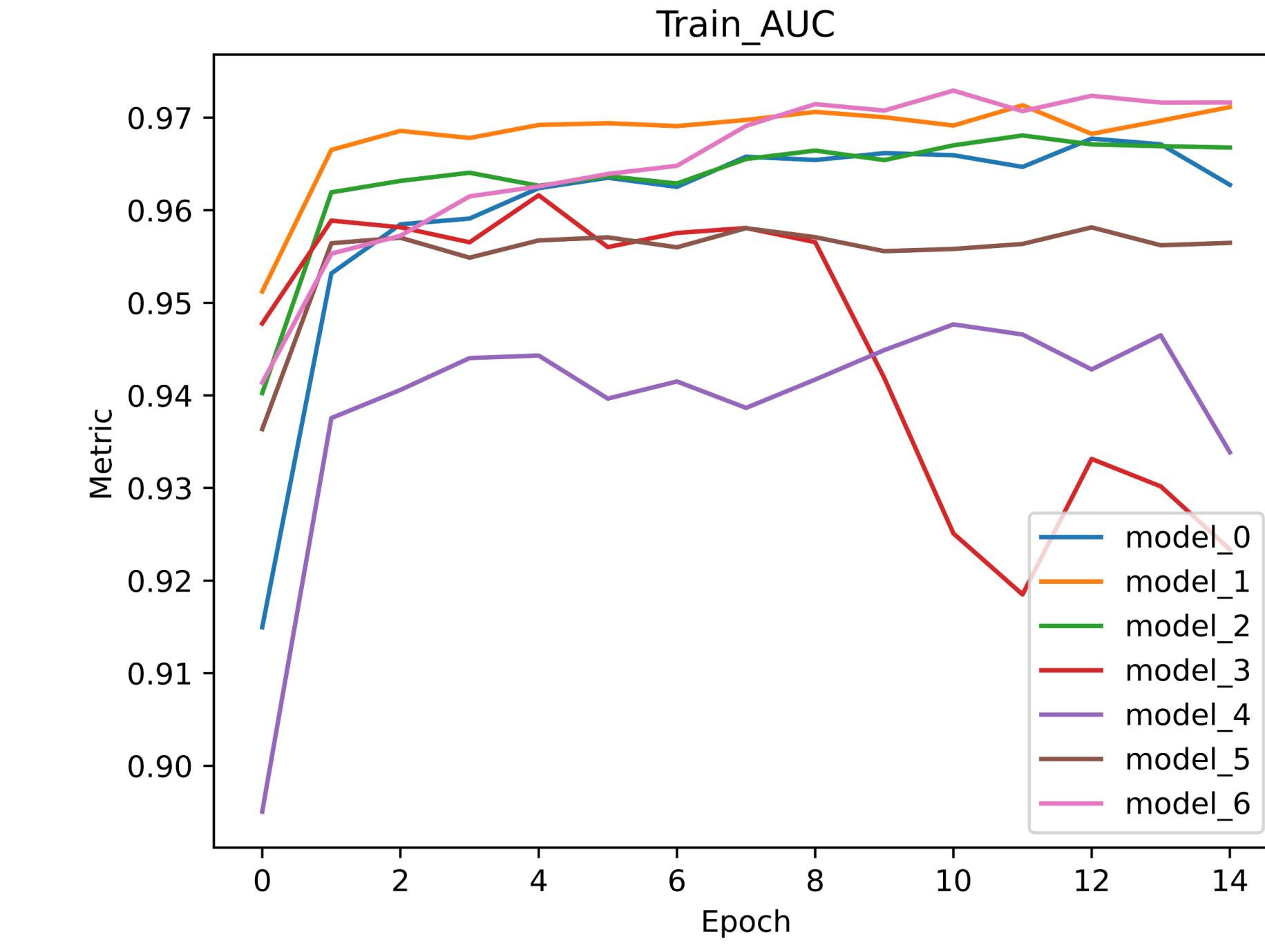
LOSS



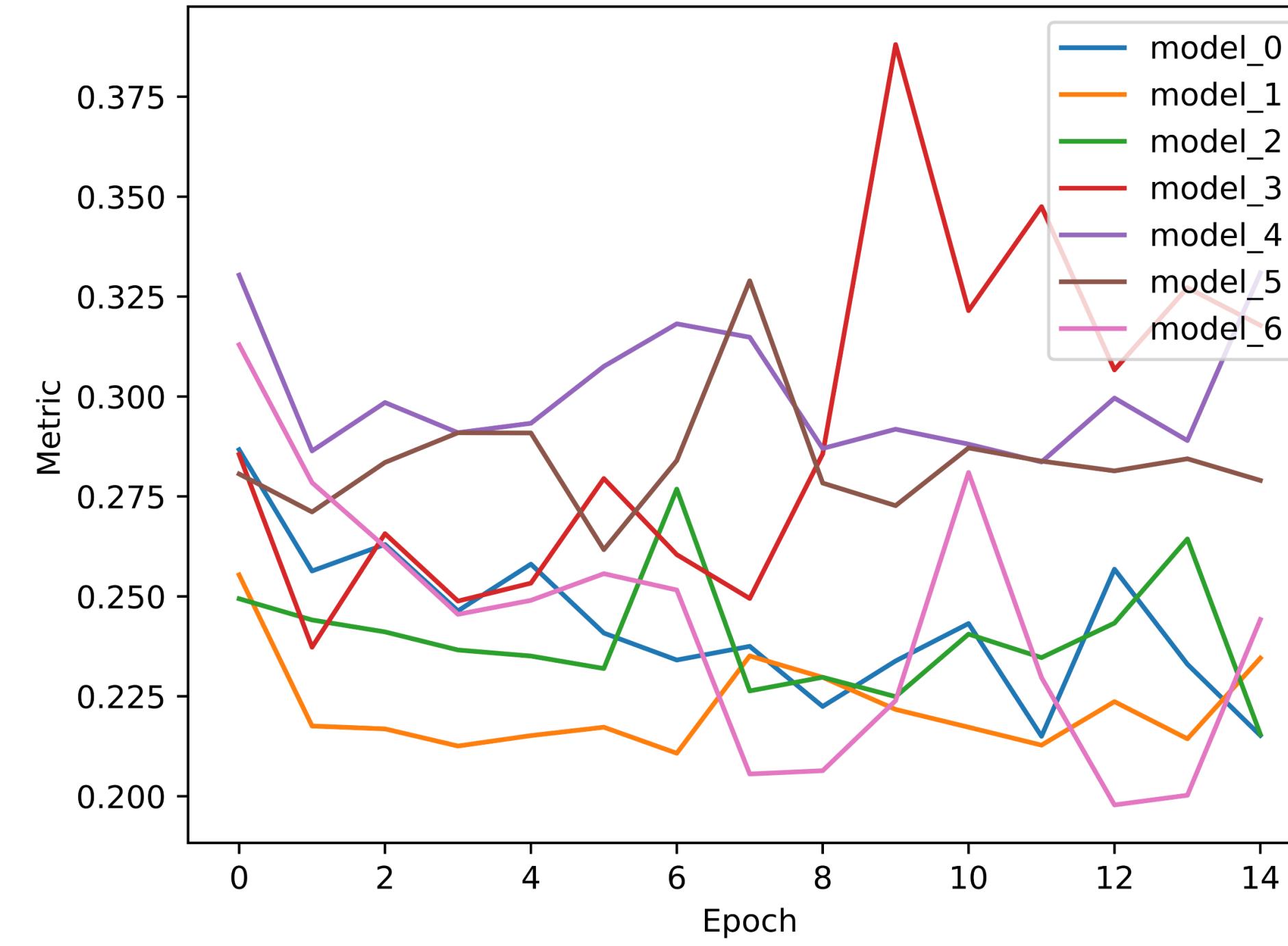
ACCURACY



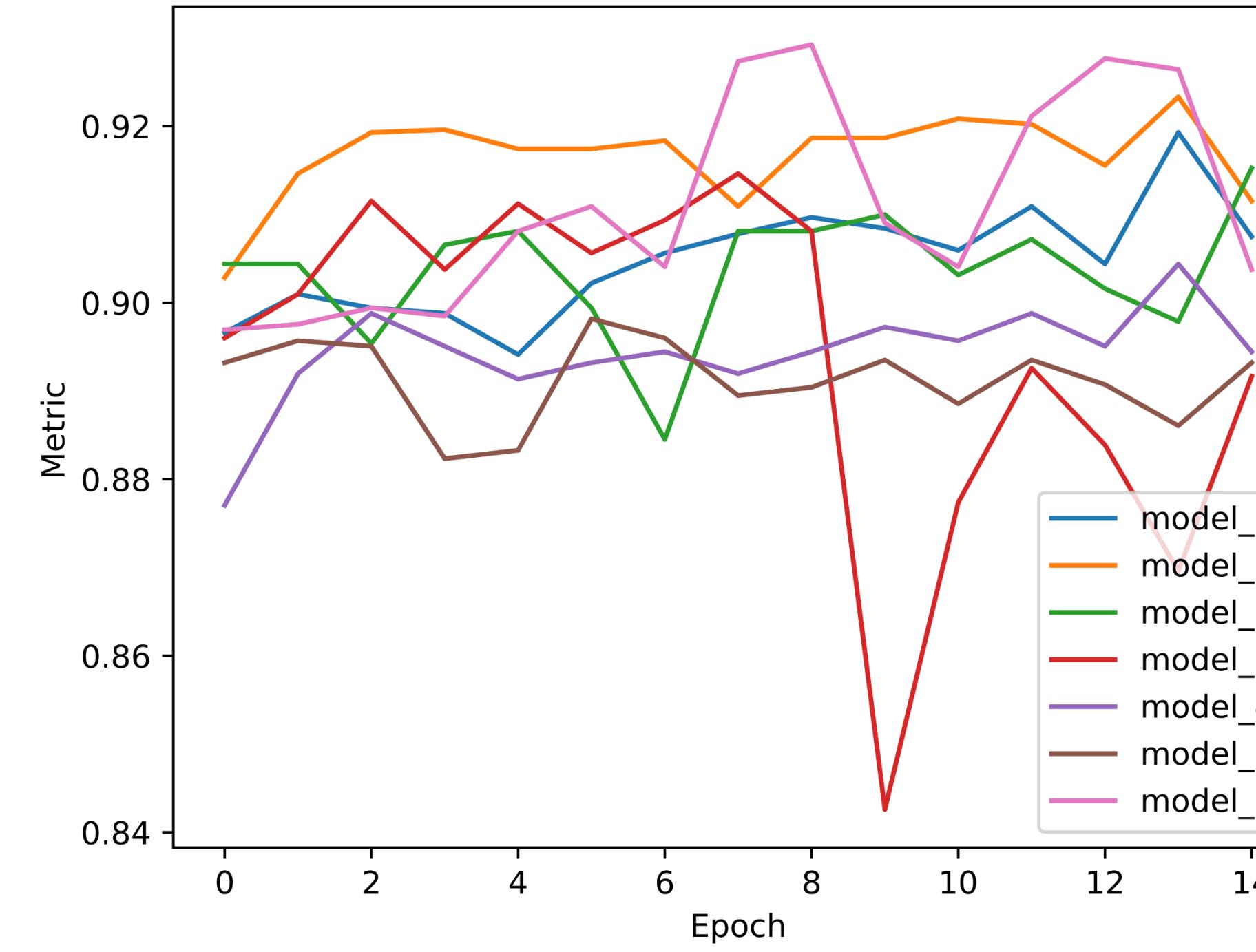
AUC



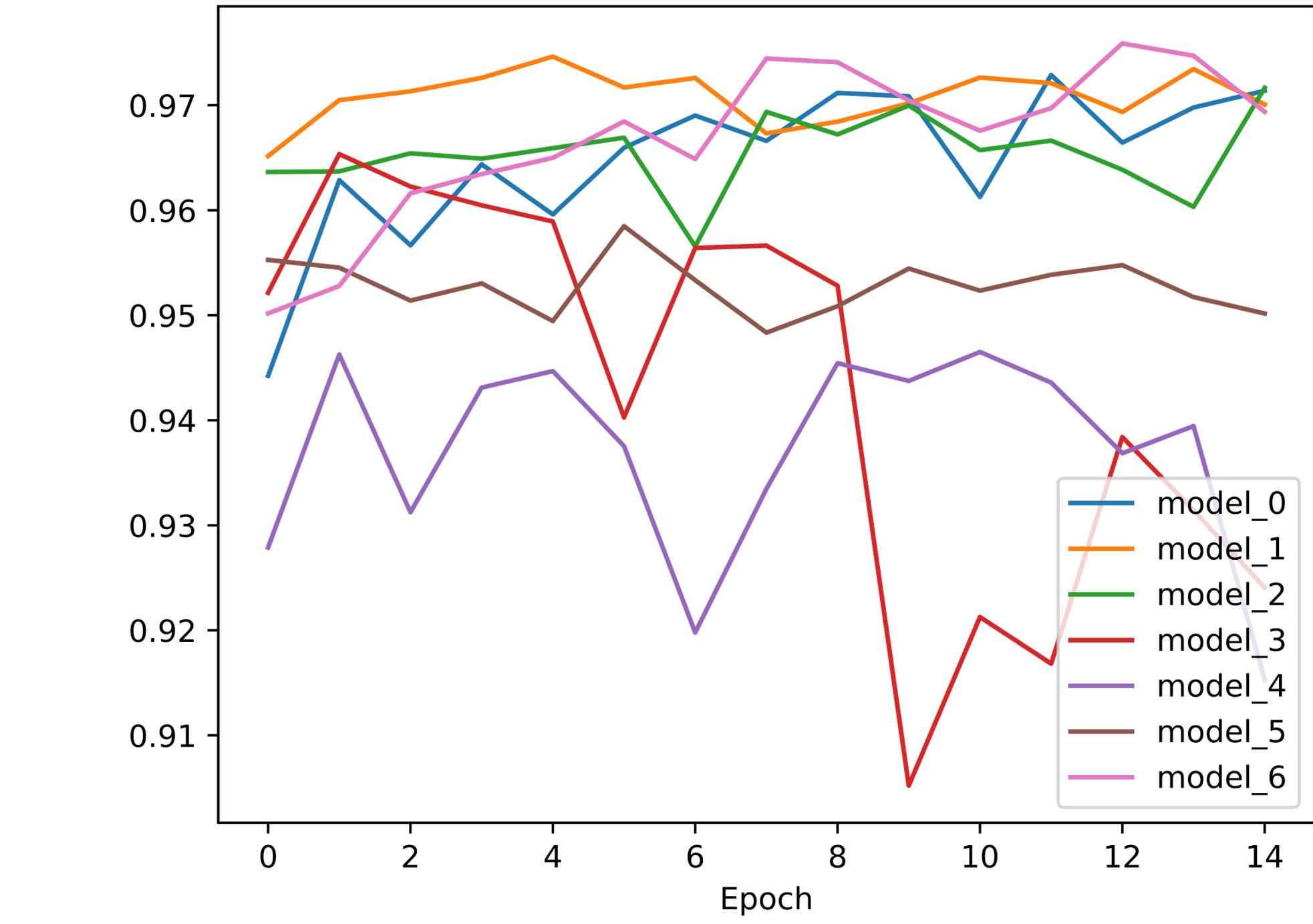
Valid_loss



Valid_accuracy

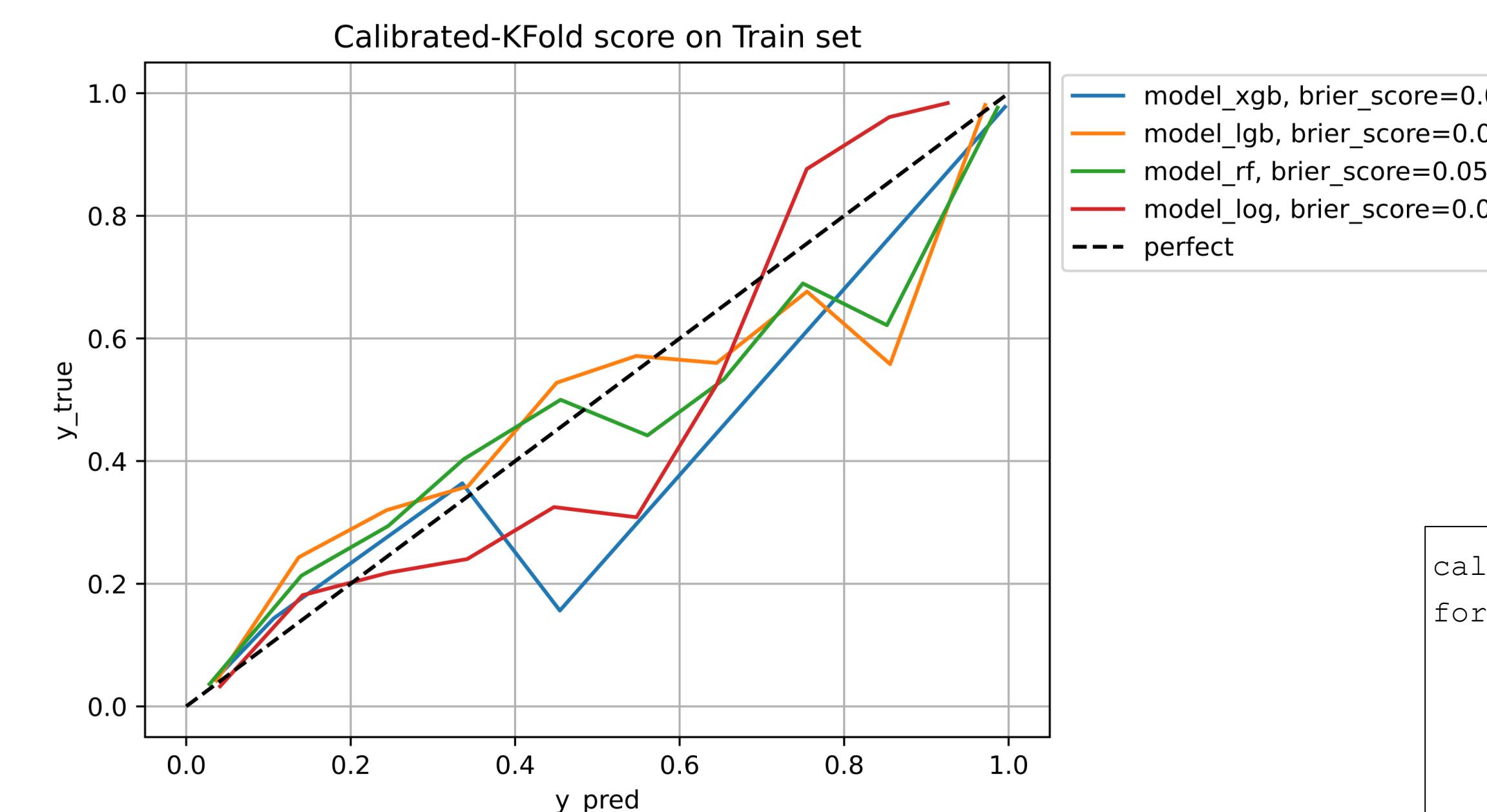
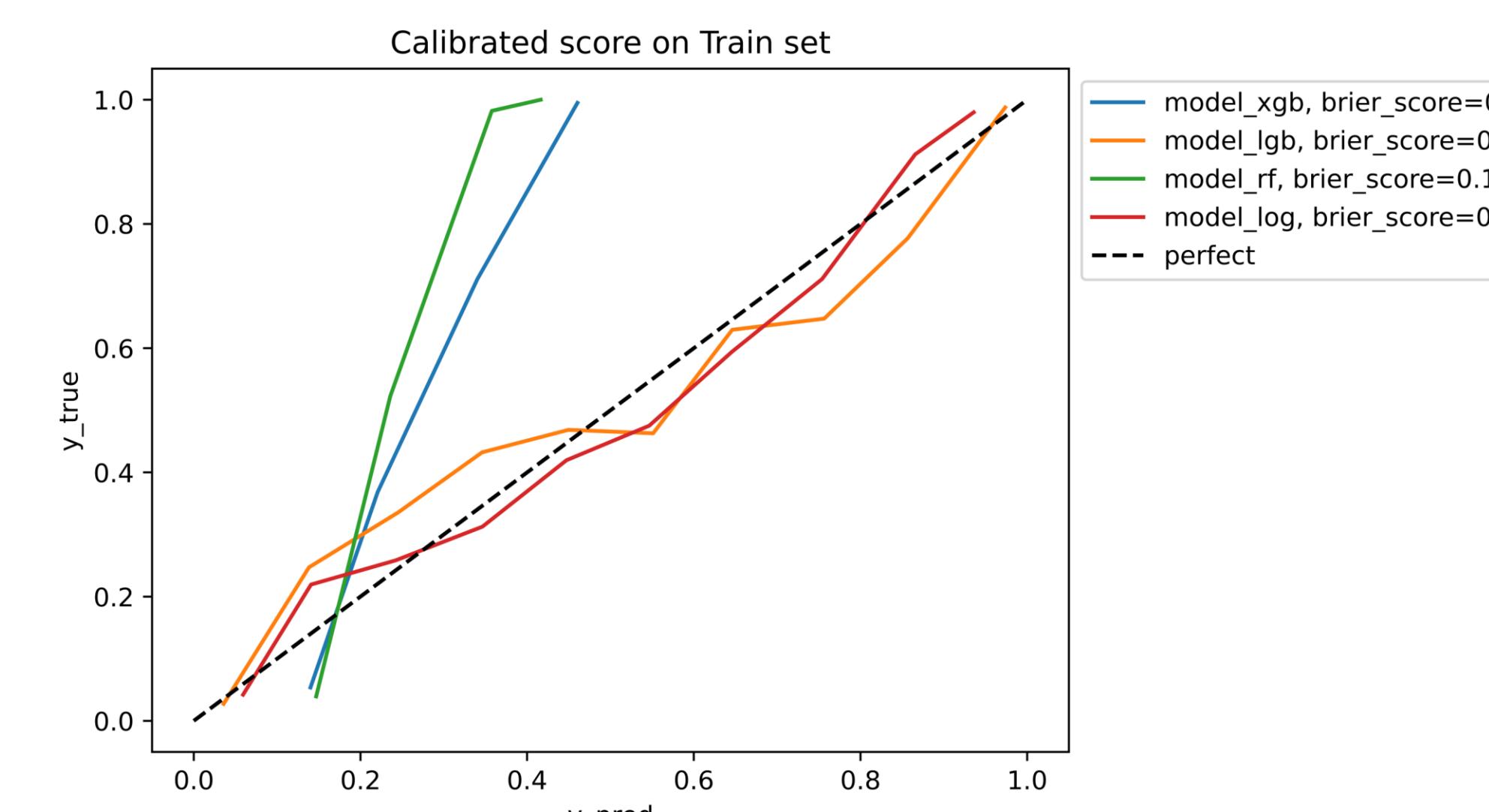
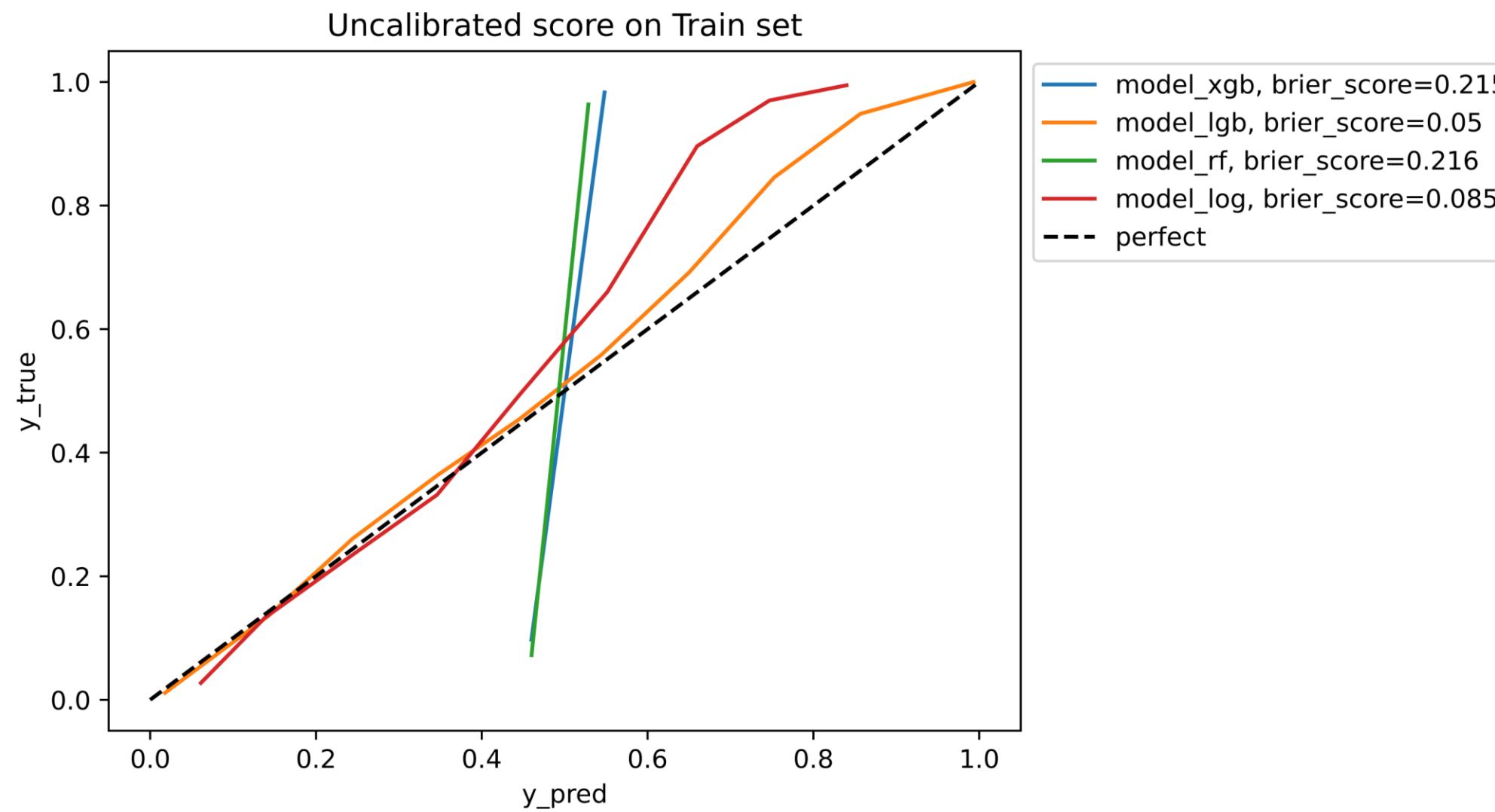


Valid_AUC

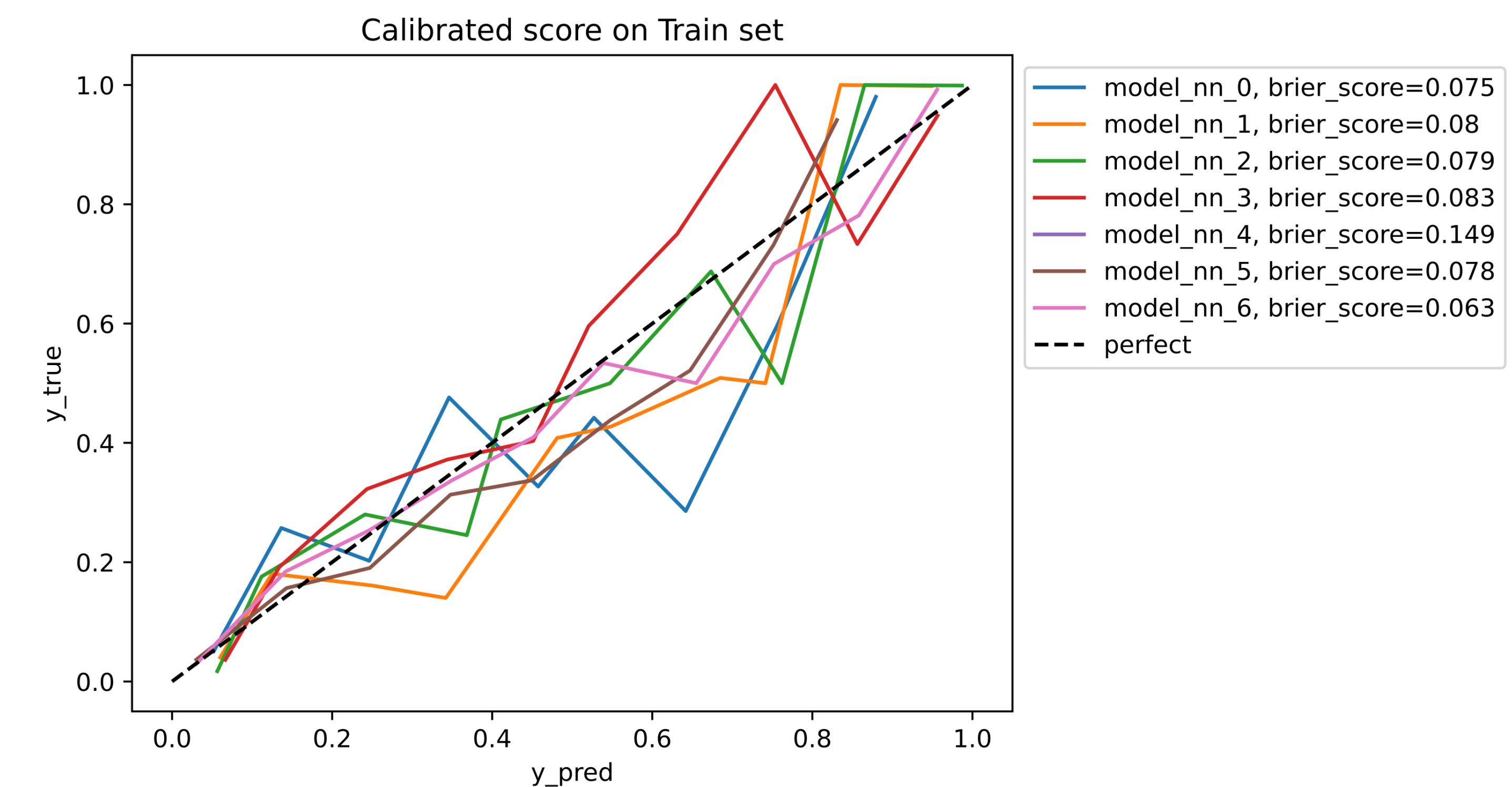
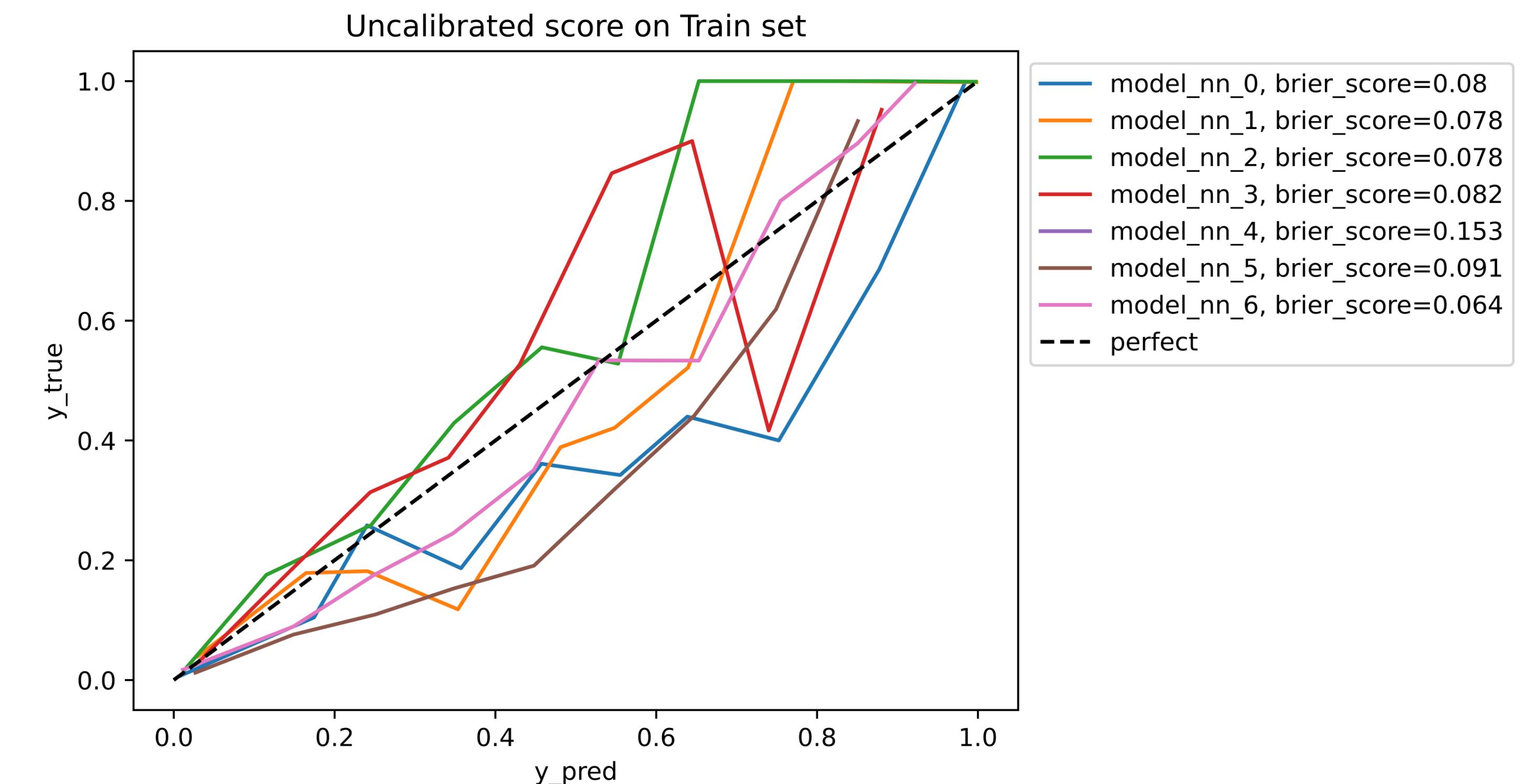


СРАВНЕНИЕ ОТКАЛИБРОВАННЫХ МОДЕЛЕЙ

Кривые надежности для классических моделей машинного обучения



Кривые надежности для нейронных сетей



```
calibrators = []
for i, model in enumerate(models_all):
    calibrator = LogisticRegression()
    calibrator.fit(pd.DataFrame(df_valid[f'{names[i]}_score']),
                   df_valid[target_col])
    calibrators.append(calibrator)
```

РЕЗУЛЬТАТЫ

Расчет вероятности ухода клиентов

Датасет должен содержать следующие поля: Client_id, Client_age, Gender, Numb_of_Prod, Salary, HasCrCard, Numb_of_years, CreditScore, Balance, IsActiveMember и быть в формате **csv**.

Загрузить файл с данными о клиентах

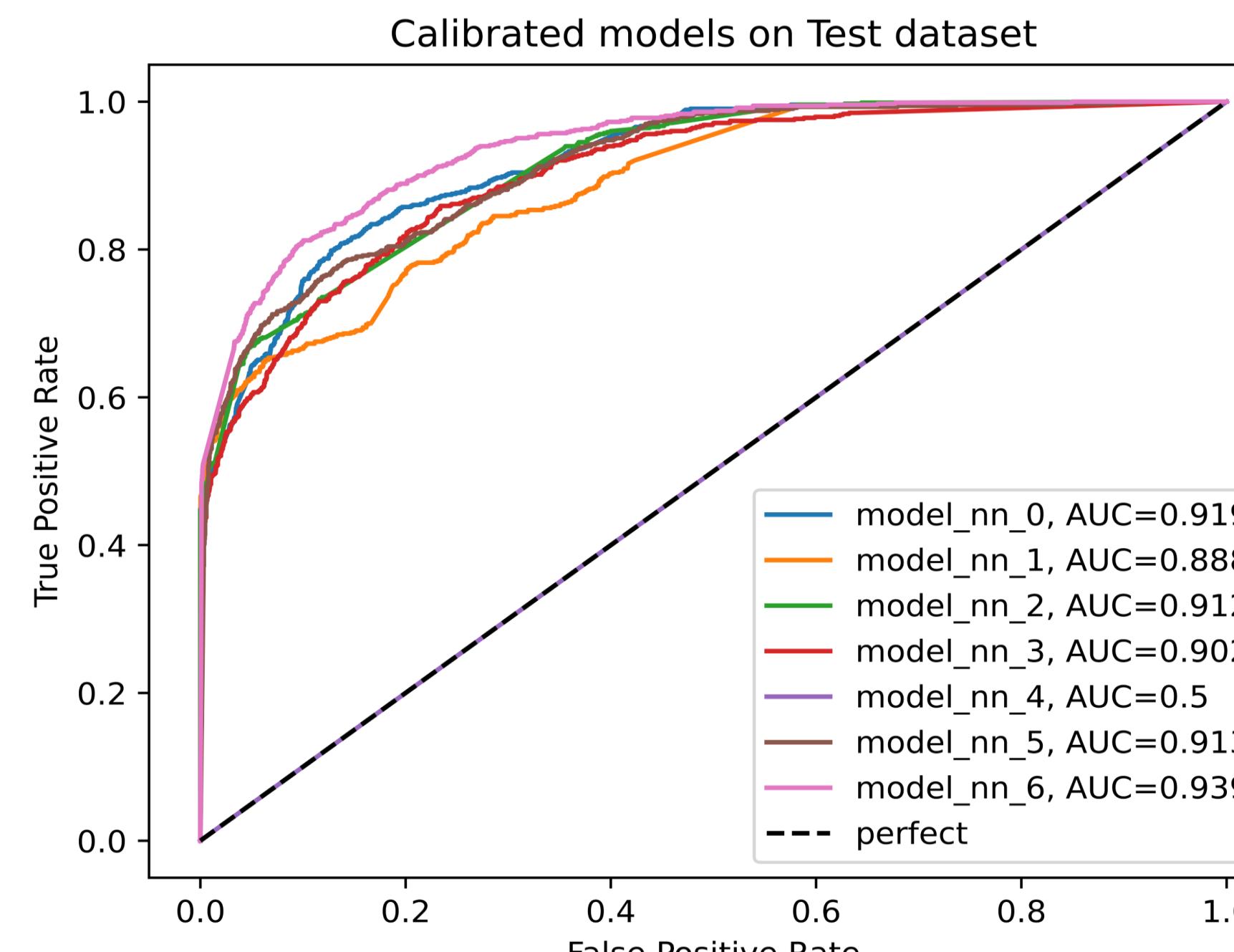
 Drag and drop file here
Limit 200MB per file Browse files

 data_for_pred.csv 216.2KB

	Unnamed: 0	Client_id	Client_age	Gender	Numb_of_Prod	Salary	HasCrCard	Num
0	5,433	15,641,575	37	Male	1	\$40K - \$60K	1	
1	2,928	15,581,198	39	Female	1	\$80K - \$120K	1	
2	5,702	708,390,633	52	Female	3	Less than \$40K	1	
3	3,062	711,502,908	44	Male	3	\$60K - \$80K	1	
4	9,199	826,061,508	49	Female	4	Less than \$40K	1	
5	9,354	15,791,501	43	Male	2	\$120K +	1	
6	5,605	15,730,272	58	Male	1	\$80K - \$120K	1	
7	9,362	807,971,658	47	Female	3	Less than \$40K	1	
8	8,196	712,121,508	43	Male	3	\$60K - \$80K	1	
9	8,868	719,157,933	51	Male	2	\$120K +	1	

Рассчитать вероятность ухода клиентов

Просмотр графиков



Отчёт по оттоку клиентов

Выберите диапазон вероятности ухода клиентов

Вероятность ухода клиентов: 0 - 20 %									
Вероятность ухода клиентов: 0 - 20 %									
Вероятность ухода клиентов: 21 - 40%									
Вероятность ухода клиентов: 40 - 60%									
Вероятность ухода клиентов: 60 - 80%									
Вероятность ухода клиентов: 80 - 100%									
4 709,106,358	-0.2609	1	0.9799	-1.1153	1	-1.1687			
5 713,061,558	0.137	1	0.059	-1.5513	1	-0.3405	0.99		
6 810,347,208	0.8334	1	1.9008	1.4175	1	-0.3405	0.99		
7 818,906,208	-1.0568	1	-1.7829	-1.1153	1	-0.7546	0.99		
8 710,930,508	-0.5594	1	0.9799	-1.1153	1	-0.3405			
9 719,661,558	0.5349	1	0.059	-0.0463	1	-0.3405	0.99		

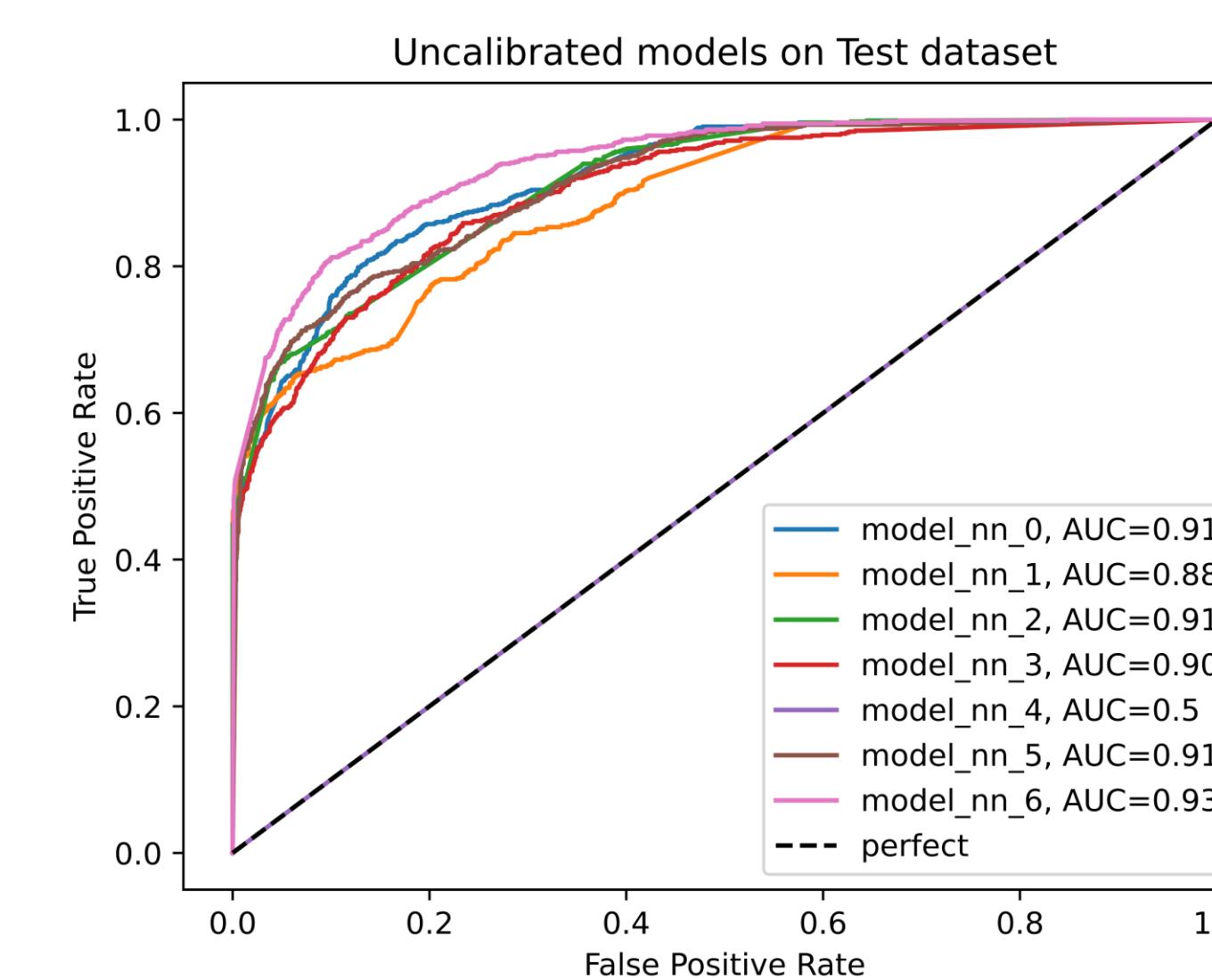
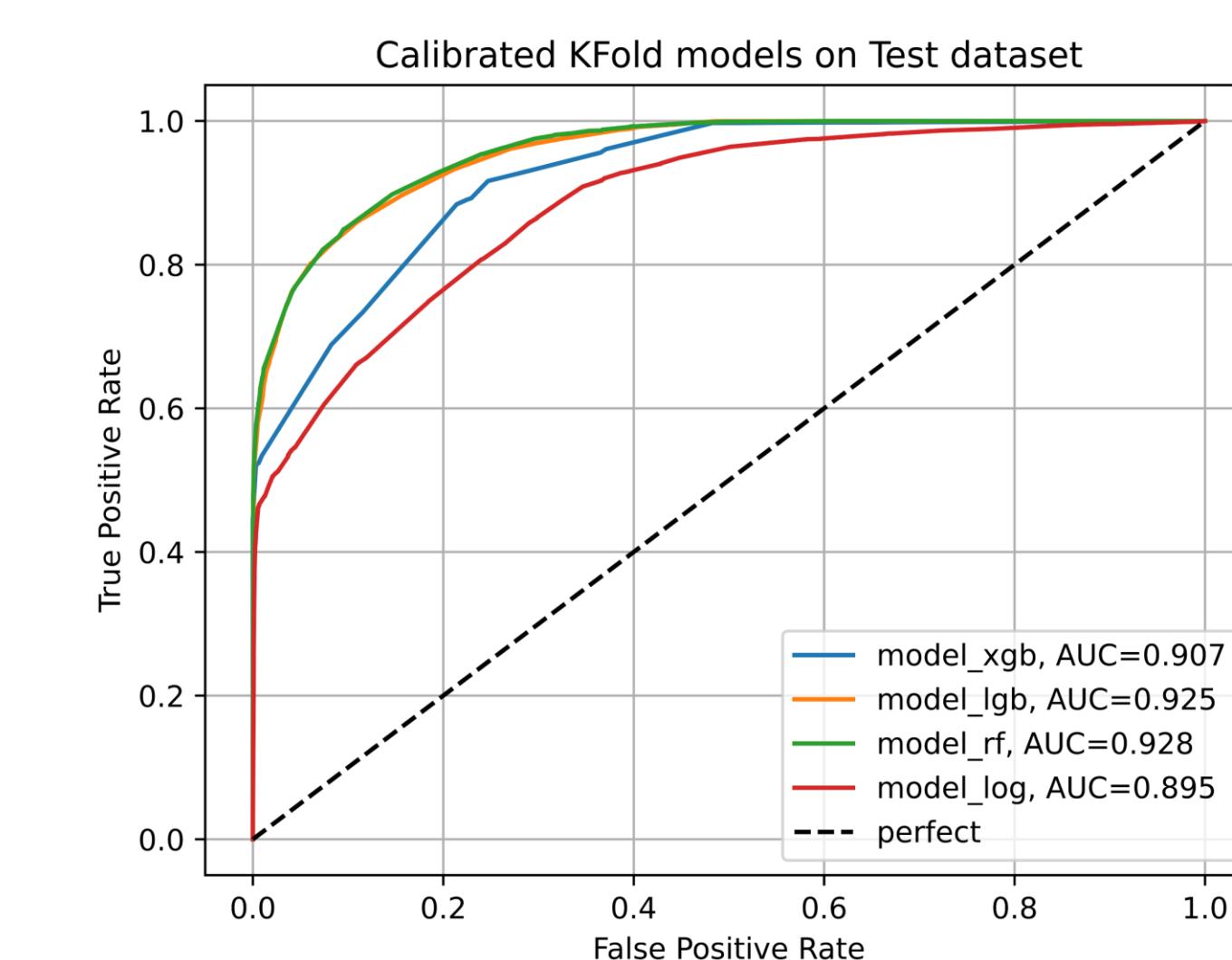
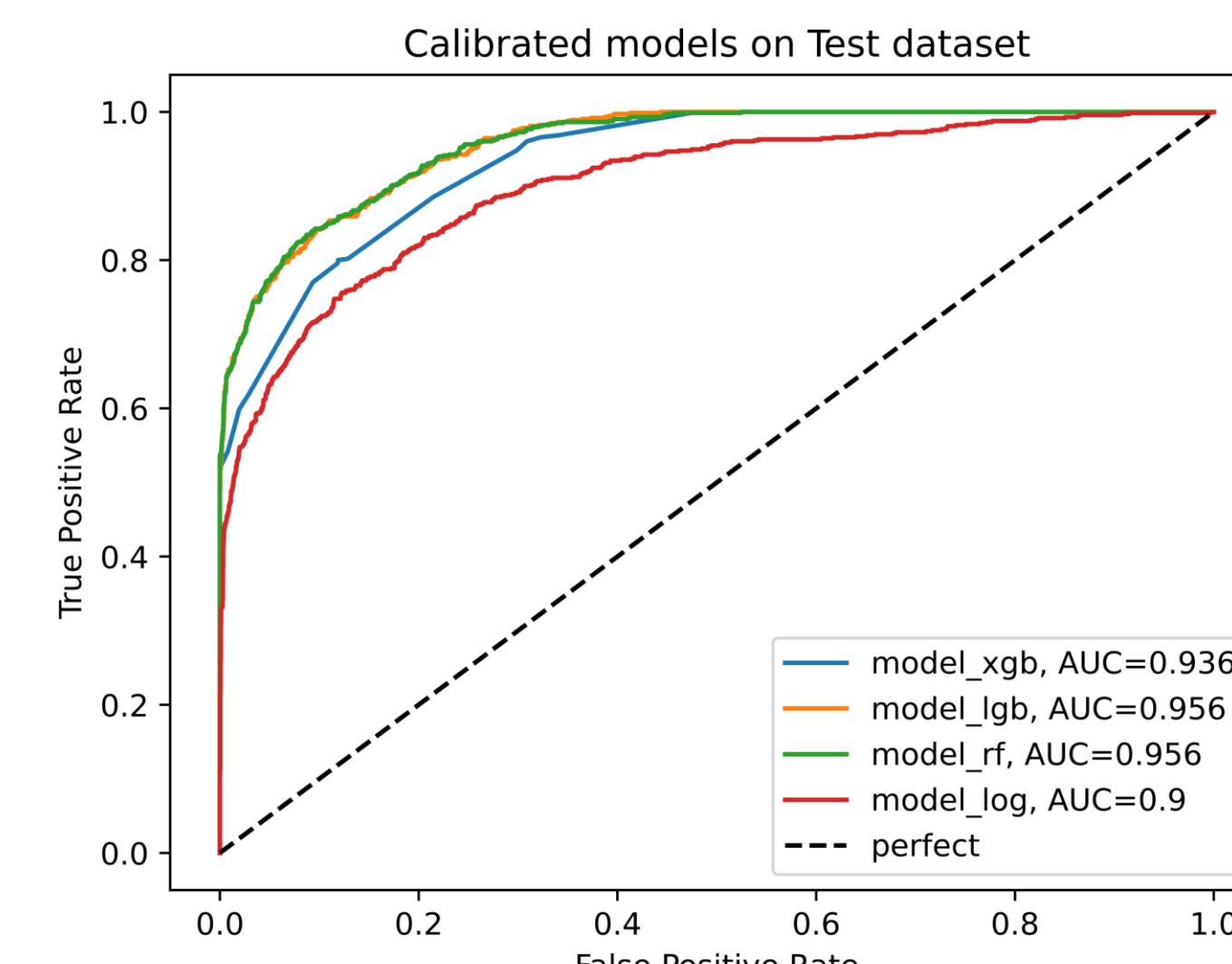
Предварительная обработка данных и расчет вероятности ухода клиента

Предсказание клиентской убыли 100

Отчёт по оттоку клиентов

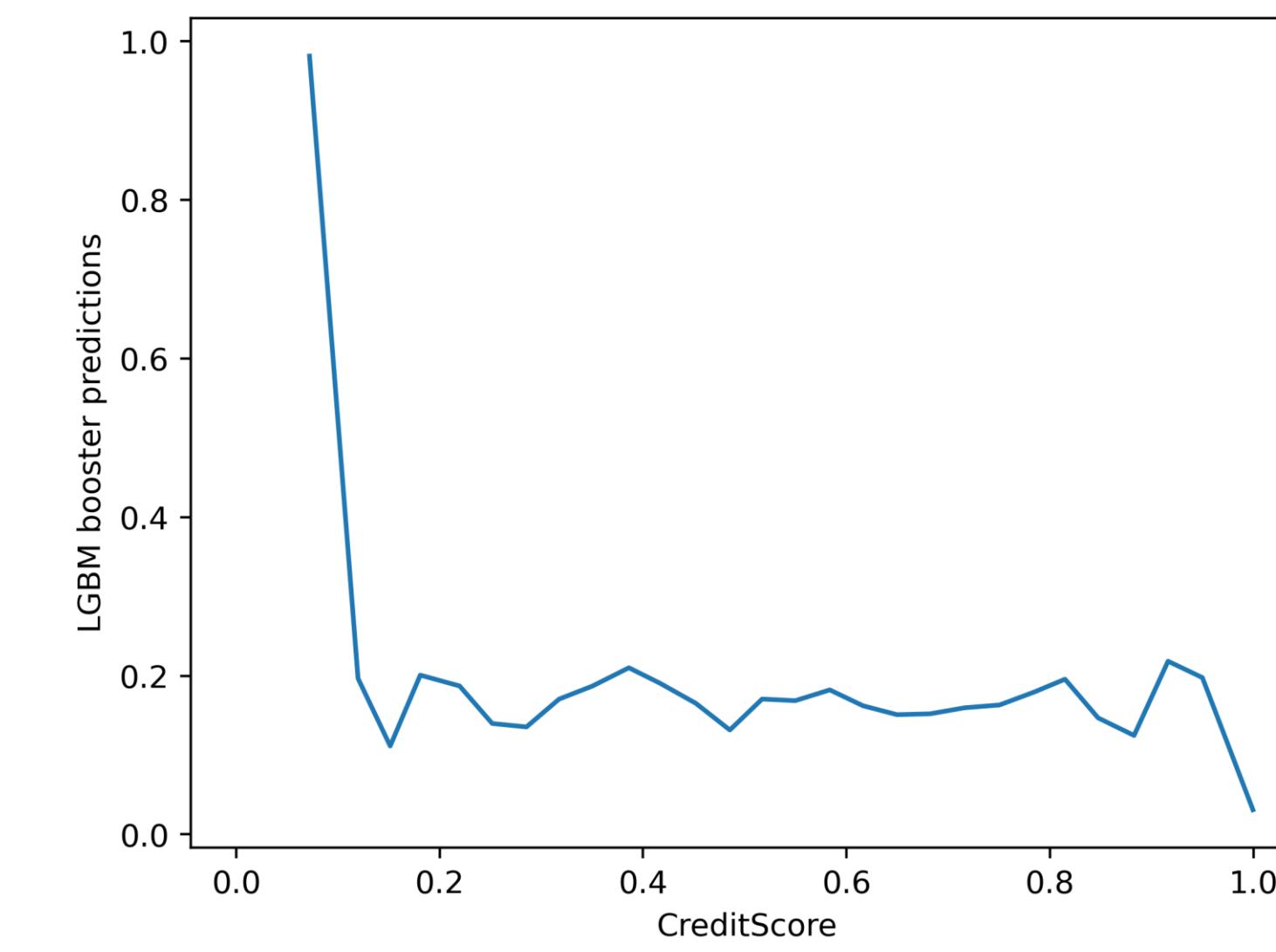
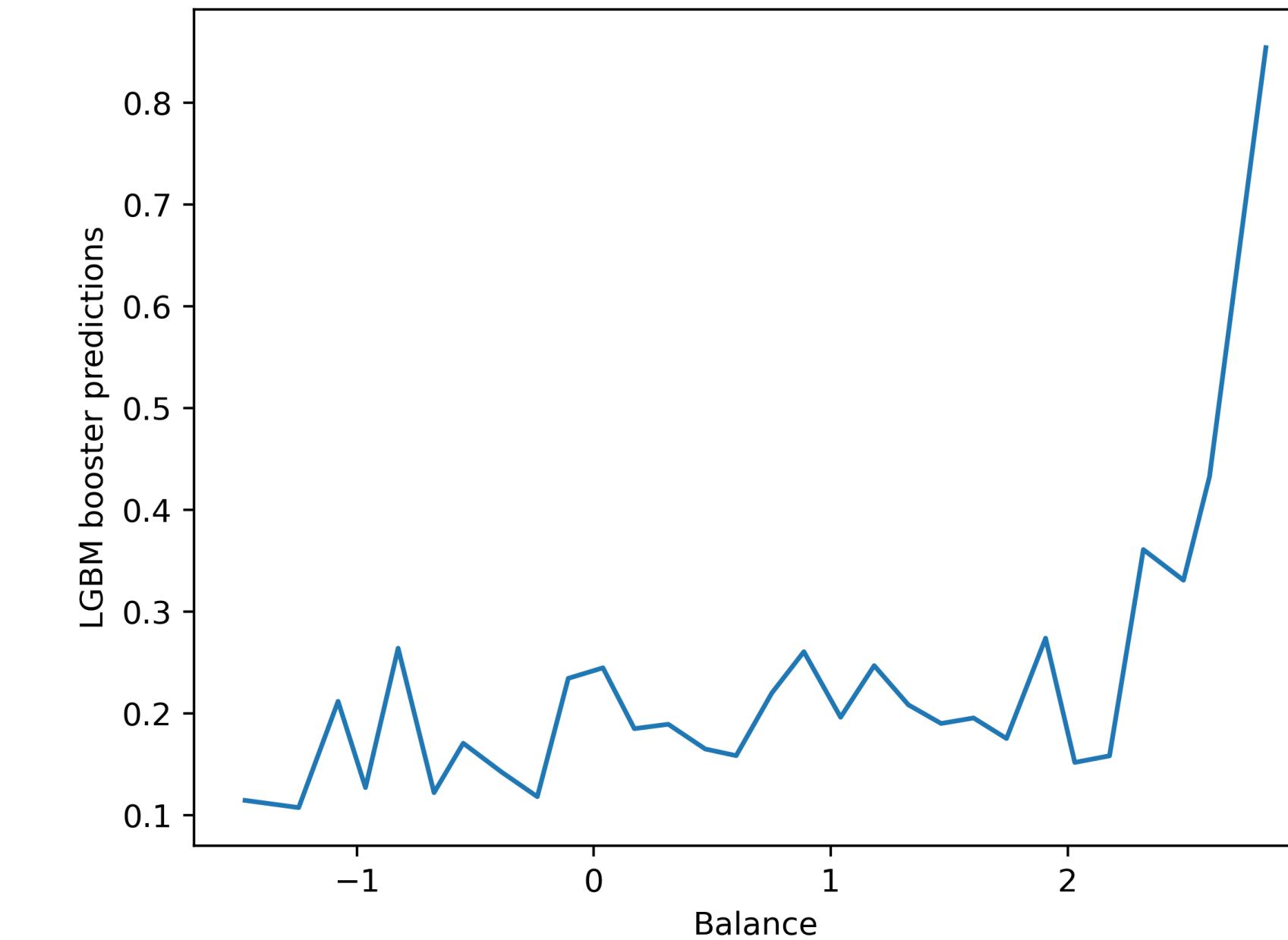
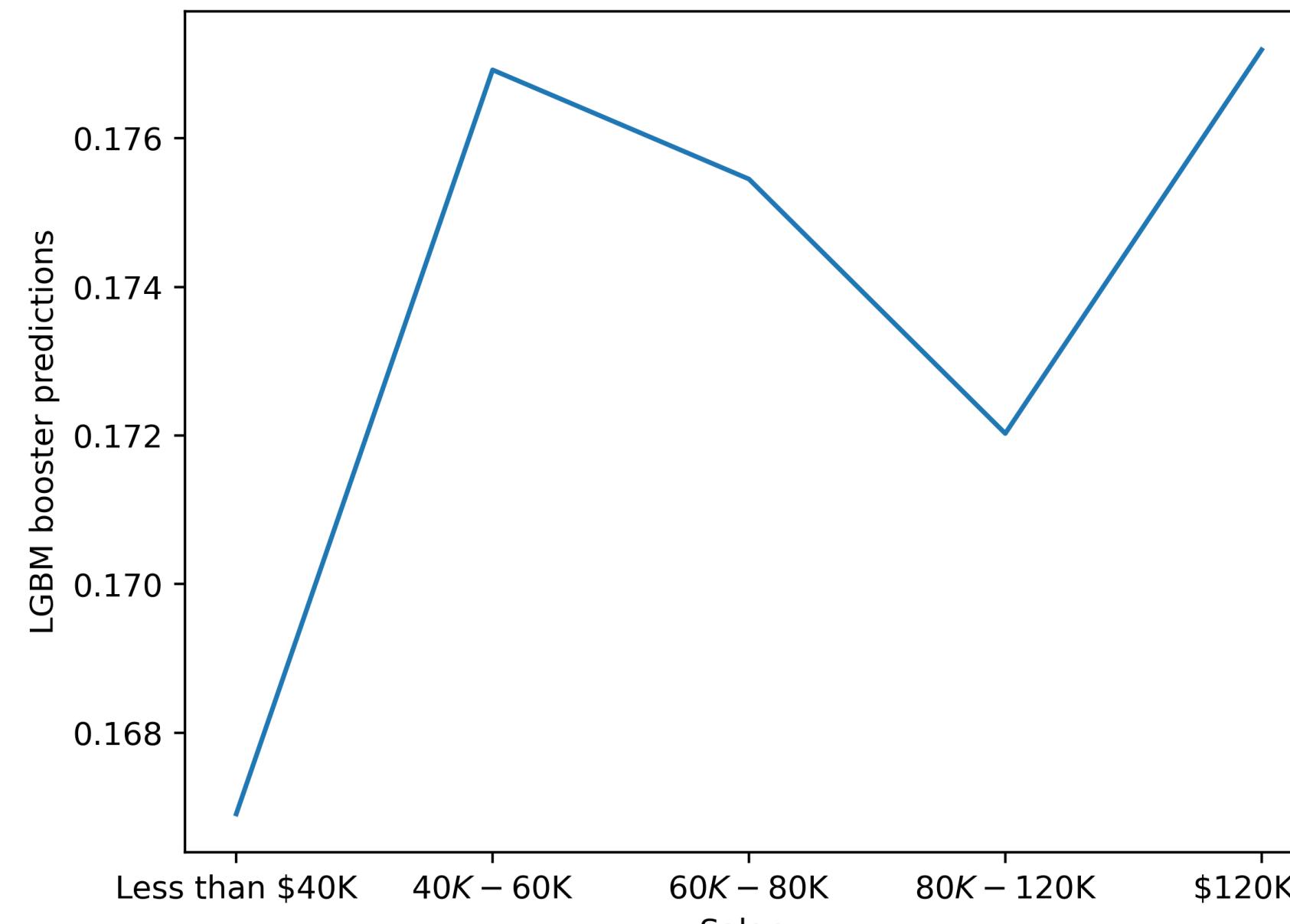
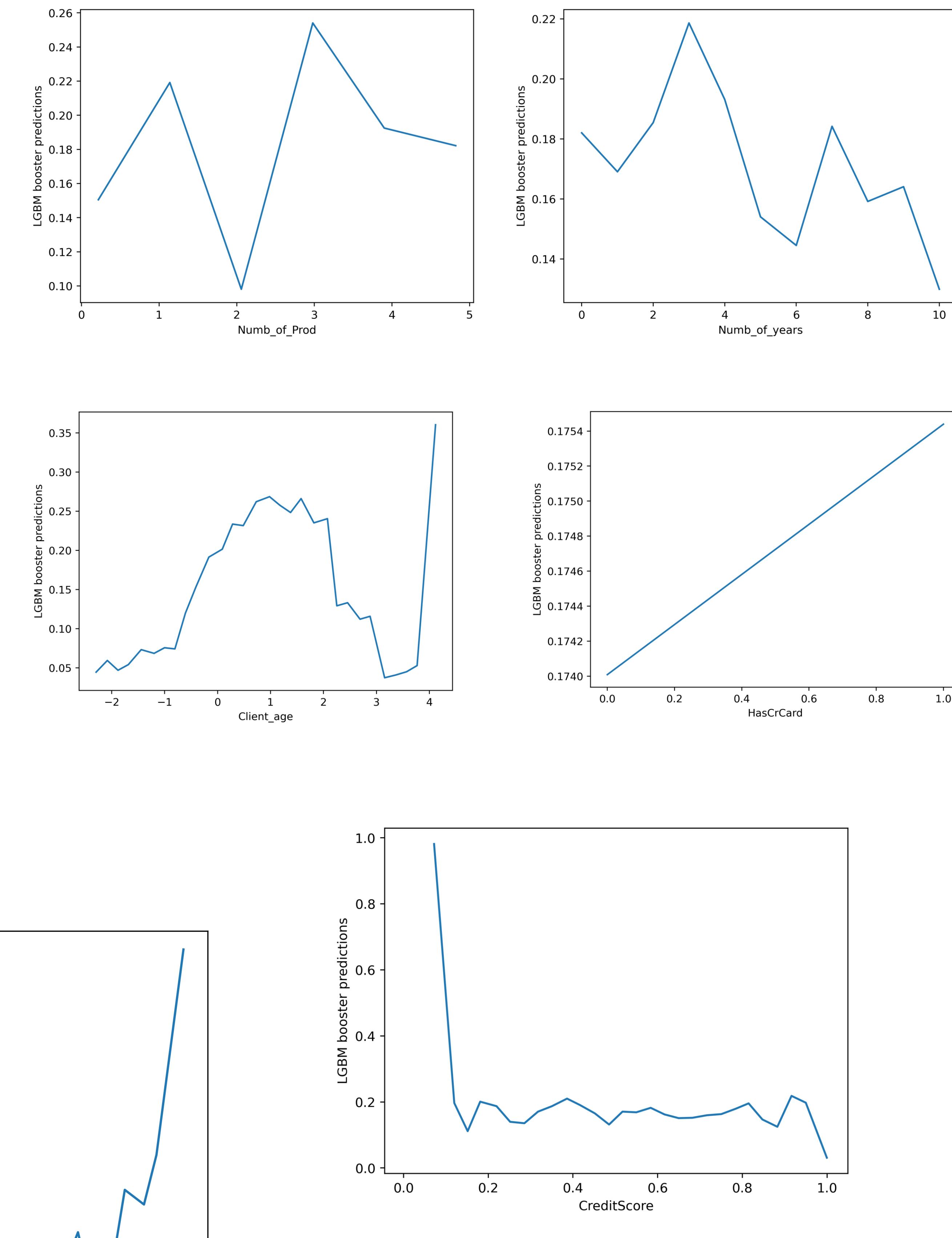
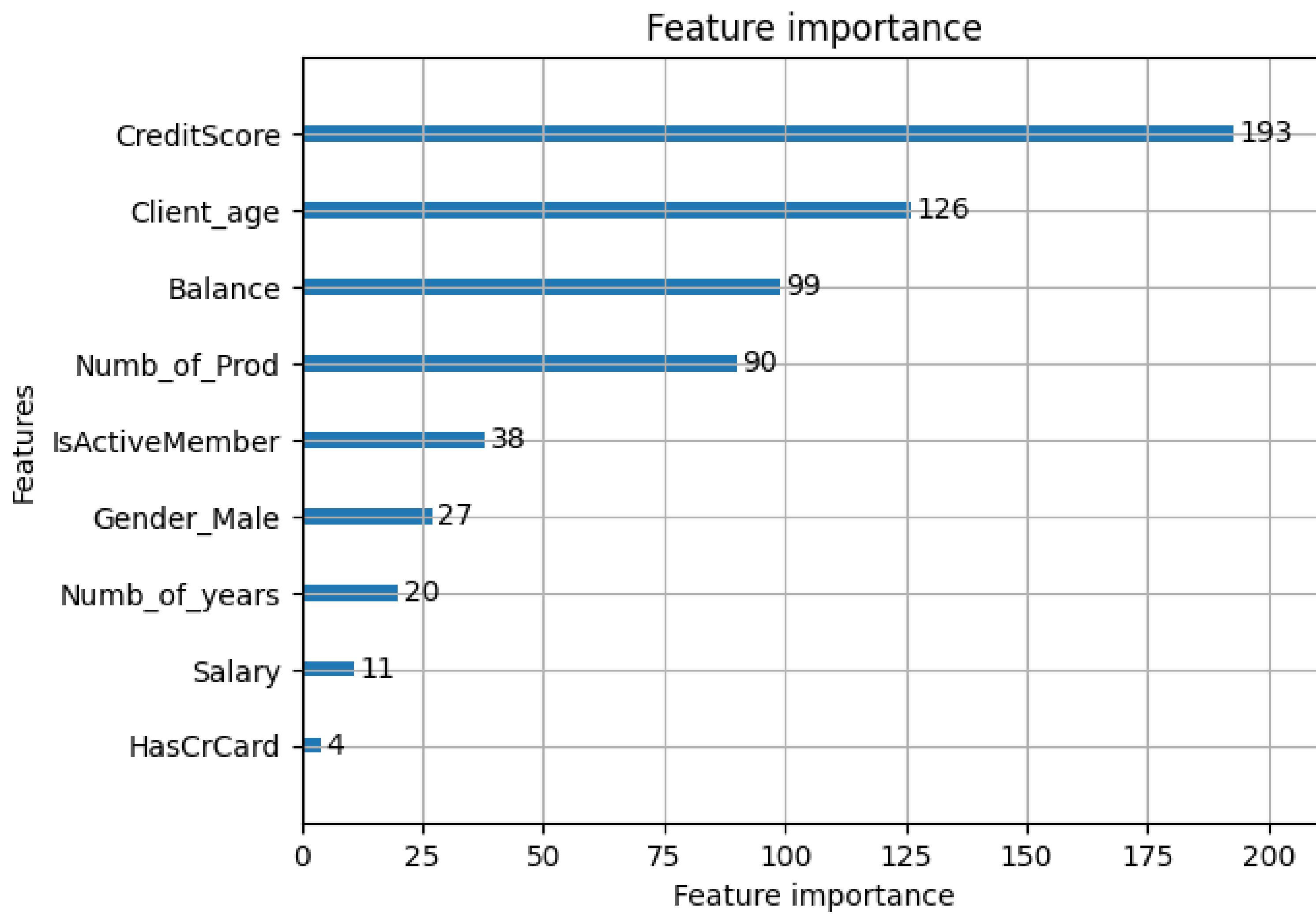
	Client_id	Client_age	Gender	Numb_of_Prod	Salary	HasCrCard	Numb_of_years	CreditScore
0	15,641,575	-0.5594	1	-0.8619	-1.5513	1	-0.7546	0.454
1	15,581,198	-0.3604	0	-0.8619	-0.0463	1	-1.5828	0.636
2	708,390,633	0.9329	0	0.9799	0.2568	1	-0.3405	0.9997
3	711,502,908	0.137	1	0.9799	-1.1153	1	-0.3405	0.9999
4	826,061,508	0.6344	0	1.9008	0.2568	1	-0.3405	0.9999
5	15,791,501	0.0375	1	0.059	1.4175	1	1.7302	0.48
6	15,730,272	1.5298	1	-0.8619	-0.0463	1	0.4878	0.538
7	807,971,658	0.4355	0	0.9799	0.2568	1	-0.3405	0.9998
8	712,121,508	0.0375	1	0.9799	-1.1153	1	-0.7546	1
9	719,157,933	0.8334	1	0.059	1.4175	1	-0.3405	0.9999

[Скачать отче](#)



РЕЗУЛЬТАТЫ

Влияние характеристик на предсказание вероятности ухода клиентов

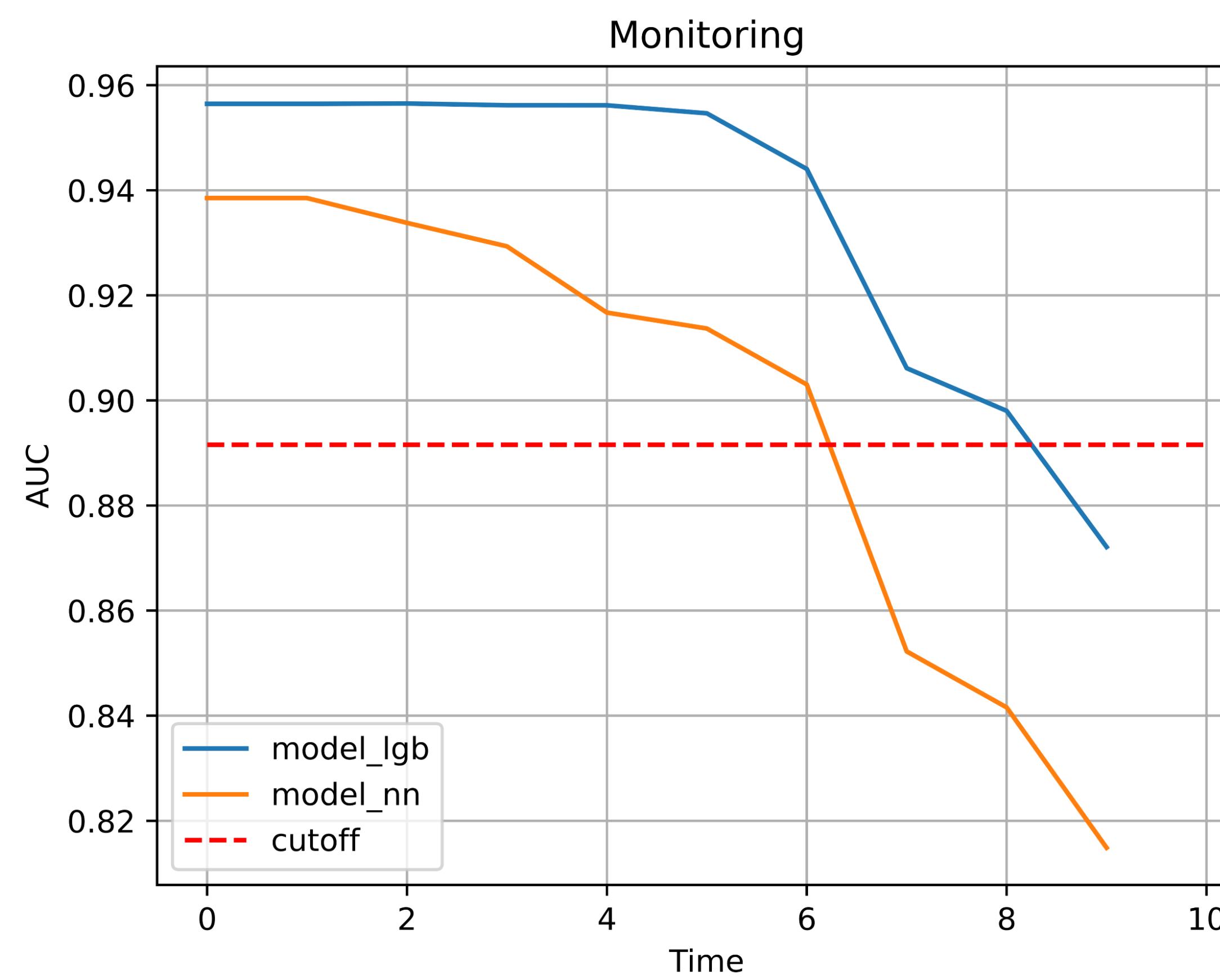


Выпускная квалификационная работа магистра					Программная подсистема расчета клиентской убыли банка						
Изм.	Лист	№ докум.	Подл.	Дата	Лит.	Масса	Масштаб				
Разраб.	Матвеенко Е.К.				Результаты						
Провер.	Пономарев А.Д.										
Н.контроль	Минигаева А.М.										

РАЗРАБОТКА ТЕХНОЛОГИИ МОНИТОРИНГА

Технология мониторинга корректности расчета клиентской убыли банка должна включать следующие шаги:

1. Запись и сохранение данных
2. Мониторинг данных
3. Разработка контрольных метрик
4. Расчет и сравнение
5. Мониторинг и анализ
6. Визуализация и отчетность
7. Реагирование и улучшение
8. Автоматизация и регулярность



Метрики построенных моделей машинного обучения

Unnamed: 0	Train_loss	Valid_loss	Test_loss	Train_accuracy	Valid_accuracy	Test_accuracy	Train_AUC	Valid_AUC	Test_AUC
model_0	0.255	0.878	0.962	0.272	0.867	0.956	0.255	0.873	0.961
model_1	0.262	0.900	0.955	0.271	0.895	0.952	0.260	0.906	0.956
model_2	0.252	0.897	0.961	0.256	0.893	0.959	0.243	0.901	0.963
model_3	0.276	0.893	0.953	0.290	0.887	0.947	0.271	0.895	0.954
model_4	0.486	0.817	0.817	0.483	0.820	0.820	0.484	0.819	0.819
model_5	0.297	0.879	0.951	0.299	0.877	0.951	0.295	0.881	0.951
model_6	0.226	0.912	0.972	0.232	0.912	0.972	0.231	0.910	0.971
model_xgb	3.229	3.256	3.151	0.910	0.910	0.913	0.758	0.750	0.761
model_lgb	2.446	2.395	2.552	0.932	0.934	0.929	0.842	0.846	0.828
model_rf	2.449	2.495	2.605	0.932	0.931	0.928	0.824	0.818	0.811
model_log	3.828	4.062	3.689	0.894	0.887	0.898	0.732	0.714	0.738
model_xgb_calibrated	6.587	6.490	6.535	0.817	0.820	0.819	0.500	0.500	0.500
model_lgb_calibrated	2.463	2.395	2.605	0.932	0.934	0.928	0.845	0.850	0.828
model_rf_calibrated	6.587	6.490	6.535	0.817	0.820	0.819	0.500	0.500	0.500
model_log_calibrated	3.795	3.995	3.608	0.895	0.889	0.900	0.755	0.743	0.763
model_nn_0_calibrated	4.231	4.577	4.432	0.883	0.873	0.877	0.821	0.798	0.810
model_nn_1_calibrated	3.604	3.782	3.393	0.900	0.895	0.906	0.763	0.748	0.771
model_nn_2_calibrated	3.719	3.849	3.563	0.897	0.893	0.901	0.735	0.722	0.742
model_nn_3_calibrated	3.786	4.062	3.680	0.895	0.887	0.898	0.729	0.711	0.736
model_nn_4_calibrated	6.587	6.490	6.535	0.817	0.820	0.819	0.500	0.500	0.500
model_nn_5_calibrated	3.630	3.547	3.366	0.899	0.902	0.907	0.790	0.793	0.802
model_nn_6_calibrated	3.171	3.156	3.223	0.912	0.912	0.911	0.826	0.826	0.821

```
# Модель стоит переобучать при ухудшении метрики на 5%
cutoff = min([auc_lgb[0], auc_nn[0]]) - min([auc_lgb[0], auc_nn[0]]) * 0.05
end = len(auc_nn)
fig = plt.figure()
plt.plot(auc_lgb, label='model_lgb')
plt.plot(auc_nn, label='model_nn')
plt.plot([0, end], [cutoff, cutoff], '--', color='red', label='cutoff')
plt.xlabel('Time')
plt.ylabel('AUC')
plt.legend(loc='lower left')
```

Выпускная квалификационная работа магистра					Программная подсистема расчета клиентской убыли банка		
Изм.	Лист	№ докум.	Подп.	Дата	Лим.	Масса	Масштаб
Разраб.	Матвеенко Е.К.						
Провер.	Пономарев А.Д.						
Н.контроль	Минихаева А.М.						

Лист 10 Листов 10
МГТУ им. Н.Э. Баумана
Факультет ИУ
Группа ИУ6-43М