Предсказание ухода сотрудников

Канаметов Азамат, Пузач Владислав, Байшев Олег

Московский физико-технический институт

25 декабря 2020 г.

1. Первый этап

2. Второй этап
   - Логистическая регрессия
   - Нейросетевые модели для классификации
   - Нейросетевые модели для временного ряда

Актуальность

Последствия ухода сотрудников:

1. нарушения дедлайнов
2. дополнительные затраты на найм сотрудников
3. временное снижение эффективности

Возможности:

1. выполнение проектов в срок
2. сохранение или повышение эффективности

Данные

- предсказываемые значения — сотрудник в конце года покинет компанию: 1, останется: 0
- соотношение классов примерно 1 к 6
- 4410 записей
- general_data.csv — общие данные о сотрудниках: возраст, отдел, уровень образования, доход и т.д
- employee_survey_data.csv — результаты первого опроса: уровни удовлетворенности рабочим пространством, удовольствия от выполнения задач, баланса между работой и личной жизнью.
- manager_survey_data.csv — результаты второго опроса: оценка сотрудниками менеджеров по их вовлеченности в работу и эффективности.

Датасет

Предобработка

- произведено one-hot кодирование категориальных признаков 'Department', 'EducationField', 'JobRole', 'MaritalStatus', 'Over18'
- бинарные признаки 'Attrition': {'No': 0, 'Yes': 1}, 'Gender': {'Female': 0, 'Male': 1}
- удалены неинформативные признаки 'EmployeeCount', 'Over18_Y', 'StandardHours': у каждого было всего одно значение для всех строк
- заполнены пропуски в столбцах 'NumCompaniesWorked', 'TotalWorkingYears', 'JobSatisfaction', 'EnvironmentSatisfaction', 'WorkLifeBalance': использовалось предсказание sklearn.ensemble.RandomForestClassifier(n_estimators=30) по известным признакам отдельно для каждого неизвестного признака
- нормализация: sklearn.preprocessing.MinMaxScaler

Поиск параметров

model	parameters	accuracy	f1 score	precision	recall
LogRegression	$C = 0.1$	$84.4_{\pm 0.4}$	$13.4_{\pm 2.4}$	$7.4_{\pm 1.4}$	$75.6_{\pm 10.5}$
LogRegression	$C = 1$	$84.2_{\pm 0.2}$	$22.6_{\pm 1.3}$	$14.1_{\pm 1.2}$	$59.2_{\pm 7.2}$
LogRegression	$C = 10$	$84.5_{\pm 0.2}$	$25.4_{\pm 1.2}$	$16.1_{\pm 1.2}$	$61.0_{\pm 6.7}$
LogRegression	$C = 100$	$84.5_{\pm 0.2}$	$25.4_{\pm 1.3}$	$16.1_{\pm 1.2}$	$61.0_{\pm 6.7}$
SVC	$C = 0.1$, $kernel = poly$	$83.7_{\pm 0.4}$	$1.3_{\pm 0.1}$	$6.9_{\pm 0.2}$	$\mathbf{100}_{\pm 0}$
SVC	$C = 1$, $kernel = poly$	$92.5_{\pm 0.8}$	$71.7_{\pm 4.5}$	$58.5_{\pm 6.4}$	$93.6_{\pm 1.7}$
SVC	$C = 10$, $kernel = poly$	$98.6_{\pm 0.5}$	$95.7_{\pm 1.5}$	$\mathbf{94.3}_{\pm 2.0}$	$97.1_{\pm 1.1}$
SVC	$C = 100$, $kernel = poly$	$\mathbf{98.8}_{\pm 0.4}$	$\mathbf{96.1}_{\pm 1.3}$	$\mathbf{95.2}_{\pm 1.9}$	$97.2_{\pm 1.0}$
DecisionTree	$max\_depth = 5$	$85.1_{\pm 0.1}$	$34.1_{\pm 1.6}$	$23.5_{\pm 1.9}$	$62.6_{\pm 6.1}$
DecisionTree	$max\_depth = 10$	$92.4_{\pm 0.7}$	$74.1_{\pm 3.1}$	$66.3_{\pm 4.8}$	$84.2_{\pm 0.8}$
DecisionTree	$max\_depth = 20$	$98.6_{\pm 0.3}$	$95.8_{\pm 0.8}$	$\mathbf{95.9}_{\pm 2.8}$	$96.0_{\pm 3.2}$
DecisionTree	$max\_depth = 40$	$97.3_{\pm 0.3}$	$94.4_{\pm 0.8}$	$\mathbf{93.9}_{\pm 1.8}$	$95.0_{\pm 1.2}$
AdaBoost	$n\_estimators = 5$	$84.0_{\pm 0.5}$	$20.7_{\pm 6.6}$	$13.2_{\pm 5.5}$	$57.6_{\pm 7.8}$
AdaBoost	$n\_estimators = 10$	$84.7_{\pm 0.2}$	$25.9_{\pm 1.7}$	$16.4_{\pm 1.7}$	$63.4_{\pm 6.9}$
AdaBoost	$n\_estimators = 20$	$84.9_{\pm 0.3}$	$31.0_{\pm 1.3}$	$20.7_{\pm 1.3}$	$62.1_{\pm 1.7}$
AdaBoost	$n\_estimators = 50$	$85.0_{\pm 0.4}$	$35.7_{\pm 2.4}$	$25.5_{\pm 3.0}$	$60.1_{\pm 3.8}$
AdaBoost	$n\_estimators = 100$	$85.3_{\pm 0.4}$	$39.7_{\pm 2.5}$	$29.5_{\pm 2.7}$	$61.0_{\pm 0.5}$
RandomForest	$n\_estimators = 5$	$97.7_{\pm 0.2}$	$92.4_{\pm 0.7}$	$87.9_{\pm 2.7}$	$97.7_{\pm 1.8}$
RandomForest	$n\_estimators = 10$	$98.3_{\pm 0.1}$	$94.6_{\pm 0.1}$	$89.9_{\pm 0.3}$	$\mathbf{100}_{\pm 0.0}$
RandomForest	$n\_estimators = 20$	$\mathbf{99.0}_{\pm 0.5}$	$\mathbf{96.7}_{\pm 1.6}$	$93.6_{\pm 3.0}$	$\mathbf{100}_{\pm 0.0}$
RandomForest	$n\_estimators = 50$	$\mathbf{99.0}_{\pm 0.3}$	$\mathbf{96.8}_{\pm 0.8}$	$\mathbf{93.8}_{\pm 1.5}$	$\mathbf{100}_{\pm 0.0}$
RandomForest	$n\_estimators = 100$	$\mathbf{99.0}_{\pm 0.3}$	$\mathbf{96.8}_{\pm 0.8}$	$93.8_{\pm 1.5}$	$\mathbf{100}_{\pm 0.0}$
GradientBoost	$n\_estimators = 10$	$84.3_{\pm 0.2}$	$8.7_{\pm 0.9}$	$4.6_{\pm 0.5}$	$95.2_{\pm 6.7}$
GradientBoost	$n\_estimators = 50$	$86.7_{\pm 0.5}$	$35.2_{\pm 3.1}$	$22.1_{\pm 2.9}$	$88.5_{\pm 6.6}$
GradientBoost	$n\_estimators = 100$	$87.7_{\pm 0.1}$	$44.1_{\pm 1.5}$	$29.5_{\pm 1.7}$	$87.7_{\pm 6.6}$
GradientBoost	$n\_estimators = 300$	$93.4_{\pm 0.2}$	$75.7_{\pm 0.5}$	$62.3_{\pm 1.3}$	$95.1_{\pm 1.9}$
GradientBoost	$n\_estimators = 400$	$95.4_{\pm 0.3}$	$84.2_{\pm 1.1}$	$74.5_{\pm 1.7}$	$97.0_{\pm 0.5}$
GradientBoost	$n\_estimators = 20$, $max\_depth = 10$	$97.8_{\pm 0.4}$	$93.1_{\pm 1.3}$	$88.9_{\pm 1.3}$	$97.7_{\pm 1.8}$
GradientBoost	$n\_estimators = 100$, $max\_depth = 10$	$\mathbf{98.9}_{\pm 0.2}$	$\mathbf{96.5}_{\pm 0.4}$	$\mathbf{93.8}_{\pm 1.5}$	$99.3_{\pm 1.0}$
GradientBoost	$n\_estimators = 300$, $max\_depth = 10$	$\mathbf{99.0}_{\pm 0.3}$	$\mathbf{96.8}_{\pm 0.8}$	$93.8_{\pm 1.5}$	$\mathbf{100}_{\pm 0.0}$

Выбранные параметры и результаты

model	parameters	f1_score
LogRegression	'C': 100	25.4
SVC	'C': 100, 'kernel': poly	96.1
DecisionTree	'max_depth': 20	95.8
AdaBoost	'n_estimators': 100	39.7
RandomForest	'n_estimators': 50	96.8
GradientBoost	'n_estimators': 300, 'max_depth': 10	96.8

Вклад

1. Канаметов Азамат: применил метод опорных векторов и композиции алгоритмов
2. Пузач Владислав: предобработка и разведочный анализ
3. Байшев Олег: нашёл датасет, оформил презентации, добавил комментарии к действиям в тетрадке

Постановка

На 2 этапе необходимо было выполнить задачи:

1. построения логистической регрессии как базового классификатора
2. определения наиболее важных признаков, исходя из результатов пункта 1
3. улучшения классификации при использовании нейросетевых моделей
4. прогнозирования временного ряда

Логистическая регрессия

Исходный датасет был предобработан на первом этапе.
Реализация: модель LogisticRegression из библиотеки sklearn
Поиск лучших параметров: GridSearchCV

Название	Значение
C	[1+10*i for i in range(11)]
solver	['lbfgs', 'liblinear']
penalty	['l2', 'none', 'l1']
max_iter	[100, 200, 300]

Лучший набор: {'C': 1, 'max_iter': 100, 'penalty': 'none', 'solver': 'lbfgs'}
f1_score на CV: 0.3026273569309039
f1_score на test: 0.2645502645502646
Длительность поиска параметров: 39.5 s
Характеристики компьютера: Intel(R) Xeon(R) CPU 2.30GHz, RAM 12.72 GB

Интерпретация

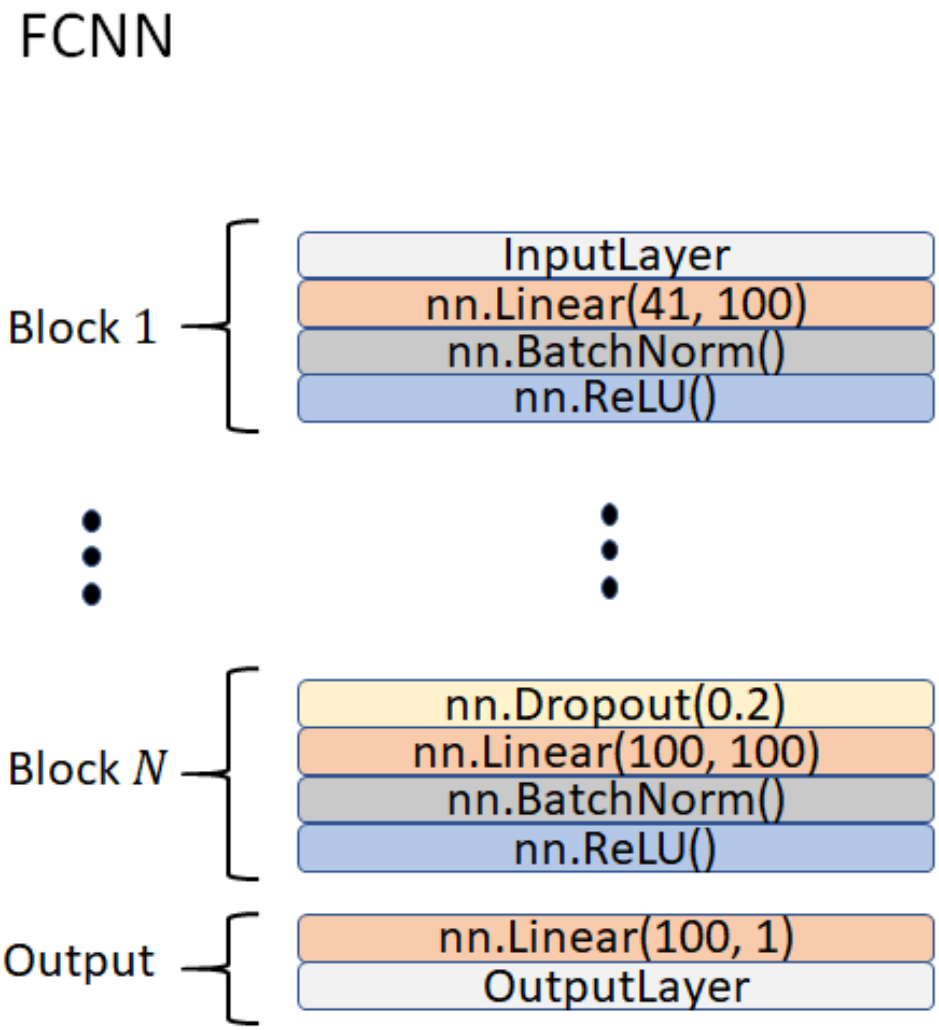
Сотрудники склонны к уходу из компании, когда:

1. долго ждут повышения
2. часто бывают в командировках
3. работали во многих компаниях
4. не состоят в браке
5. учились на HR

Люди скорее всего останутся в компании на следующий год, если они:

1. имеют большой трудовой стаж
2. долгое время работают с текущим менеджером
3. солидного возраста
4. довольны работой
5. проходили обучение в течение последнего года

Архитектура



Параметры и результаты

model	parameters	accuracy	f1 score	precision	recall
MLP	$max\_iter = 500$, $hidden\_layer\_sizes = (50, 1)$	$88.1_{\pm 5.4}$	$53.7_{\pm 22.2}$	$45.2_{\pm 22.3}$	$67.9_{\pm 19.5}$
MLP	$max\_iter = 1000$, $hidden\_layer\_sizes = (50, 1)$	$89.0_{\pm 6.7}$	$57.1_{\pm 26.9}$	$49.9_{\pm 28.9}$	$69.5_{\pm 21.6}$
MLP	$max\_iter = 500$, $hidden\_layer\_sizes = (50, 2)$	$92.1_{\pm 4.9}$	$71.7_{\pm 18.9}$	$64.2_{\pm 20.9}$	$82.4_{\pm 15.3}$
MLP	$max\_iter = 1000$, $hidden\_layer\_sizes = (50, 2)$	$94.2_{\pm 6.3}$	$79.0_{\pm 23.8}$	$75.3_{\pm 27.5}$	$84.8_{\pm 17.2}$
MLP	$max\_iter = 500$, $hidden\_layer\_sizes = (50, 3)$	$\mathbf{97.2}_{\pm 1.5}$	$\mathbf{91.2}_{\pm 4.8}$	$\mathbf{88.5}_{\pm 6.6}$	$\mathbf{94.3}_{\pm 3.0}$
MLP	$max\_iter = 1000$, $hidden\_layer\_sizes = (50, 3)$	$\mathbf{98.4}_{\pm 0.7}$	$\mathbf{94.9}_{\pm 2.1}$	$\mathbf{94.0}_{\pm 2.7}$	$\mathbf{95.8}_{\pm 1.7}$
MLP	$max\_iter = 500$, $hidden\_layer\_sizes = (100, 1)$	$85.9_{\pm 3.5}$	$51.5_{\pm 11.1}$	$45.4_{\pm 9.0}$	$59.5_{\pm 14.5}$
MLP	$max\_iter = 1000$, $hidden\_layer\_sizes = (100, 1)$	$86.4_{\pm 4.1}$	$53.0_{\pm 13.3}$	$46.6_{\pm 10.6}$	$61.7_{\pm 17.5}$
MLP	$max\_iter = 500$, $hidden\_layer\_sizes = (100, 2)$	$90.6_{\pm 4.8}$	$64.2_{\pm 20.6}$	$55.8_{\pm 21.8}$	$77.9_{\pm 20.4}$
MLP	$max\_iter = 1000$, $hidden\_layer\_sizes = (100, 2)$	$91.6_{\pm 5.9}$	$67.5_{\pm 24.0}$	$59.0_{\pm 25.7}$	$81.2_{\pm 22.5}$
MLP	$max\_iter = 500$, $hidden\_layer\_sizes = (100, 3)$	$96.8_{\pm 0.3}$	$89.8_{\pm 1.3}$	$86.1_{\pm 2.4}$	$94.0_{\pm 0.4}$
MLP	$max\_iter = 1000$, $hidden\_layer\_sizes = (100, 3)$	$\mathbf{98.5}_{\pm 0.5}$	$\mathbf{95.2}_{\pm 1.4}$	$\mathbf{94.0}_{\pm 2.2}$	$\mathbf{96.5}_{\pm 1.9}$

Рис.: MLP

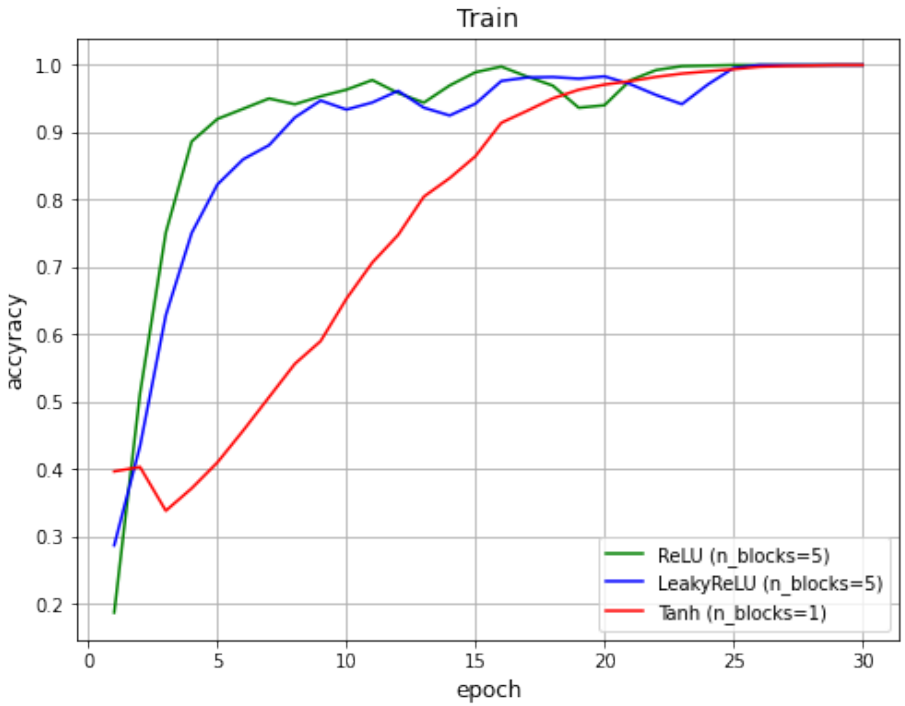
Параметры и результаты 2

model	block structure	N blocks	N parameters	accuracy	f1 score	precision	recall
FCNN	Linear	0	42	$84.1_{\pm 0.6}$	$8.3_{\pm 2.7}$	$4.4_{\pm 1.5}$	$80.2_{\pm 4.4}$
FCNN	Linear/Bn/ReLU	1	4501	$97.0_{\pm 1.1}$	$90.2_{\pm 3.9}$	$85.0_{\pm 5.3}$	$96.3_{\pm 2.2}$
FCNN	Linear/Bn/ReLU	2	14801	$98.9_{\pm 0.6}$	$96.8_{\pm 1.9}$	$97.9_{\pm 1.5}$	$95.7_{\pm 2.3}$
FCNN	Linear/Bn/ReLU	3	25101	$99.0_{\pm 0.4}$	$96.9_{\pm 1.4}$	$96.0_{\pm 2.1}$	$97.9_{\pm 1.2}$
FCNN	Linear/Bn/ReLU	4	35401	$99.4_{\pm 0.5}$	$98.3_{\pm 1.5}$	$\mathbf{98.6}_{\pm 1.5}$	$97.9_{\pm 1.5}$
FCNN	Linear/Bn/ReLU	5	45701	$\mathbf{99.7}_{\pm 0.2}$	$\mathbf{98.9}_{\pm 0.8}$	$\mathbf{99.1}_{\pm 0.9}$	$\mathbf{98.8}_{\pm 0.9}$
FCNN	Linear/Bn/LeakyReLU	1	4501	$94.9_{\pm 1.3}$	$82.6_{\pm 5.6}$	$75.2_{\pm 8.2}$	$92.0_{\pm 1.4}$
FCNN	Linear/Bn/LeakyReLU	2	14801	$97.3_{\pm 1.2}$	$91.4_{\pm 4.3}$	$88.6_{\pm 6.0}$	$94.4_{\pm 2.4}$
FCNN	Linear/Bn/LeakyReLU	3	25101	$98.3_{\pm 0.9}$	$94.6_{\pm 3.1}$	$91.4_{\pm 4.6}$	$98.2_{\pm 1.7}$
FCNN	Linear/Bn/LeakyReLU	4	35401	$98.5_{\pm 0.6}$	$95.4_{\pm 0.2}$	$95.8_{\pm 3.1}$	$94.9_{\pm 1.1}$
FCNN	Linear/Bn/LeakyReLU	5	45701	$\mathbf{99.5}_{\pm 0.3}$	$\mathbf{98.5}_{\pm 1.0}$	$\mathbf{98.1}_{\pm 1.2}$	$98.8_{\pm 0.9}$
FCNN	Linear/Bn/Tanh	1	4501	$99.4_{\pm 0.5}$	$98.1_{\pm 1.5}$	$97.4_{\pm 2.3}$	$\mathbf{98.8}_{\pm 0.9}$
FCNN	Linear/Bn/Tanh	2	14801	$98.6_{\pm 0.4}$	$95.4_{\pm 1.3}$	$91.2_{\pm 2.3}$	$\mathbf{100}_{\pm 0.0}$
FCNN	Linear/Bn/Tanh	3	25101	$99.0_{\pm 0.4}$	$97.0_{\pm 1.3}$	$97.5_{\pm 0.7}$	$96.6_{\pm 2.0}$
FCNN	Linear/Bn/Tanh	4	35401	$98.6_{\pm 0.7}$	$95.7_{\pm 2.3}$	$96.5_{\pm 3.2}$	$95.0_{\pm 1.6}$
FCNN	Linear/Bn/Tanh	5	45701	$\mathbf{99.5}_{\pm 0.3}$	$\mathbf{98.5}_{\pm 1.0}$	$98.1_{\pm 1.2}$	$98.8_{\pm 0.9}$

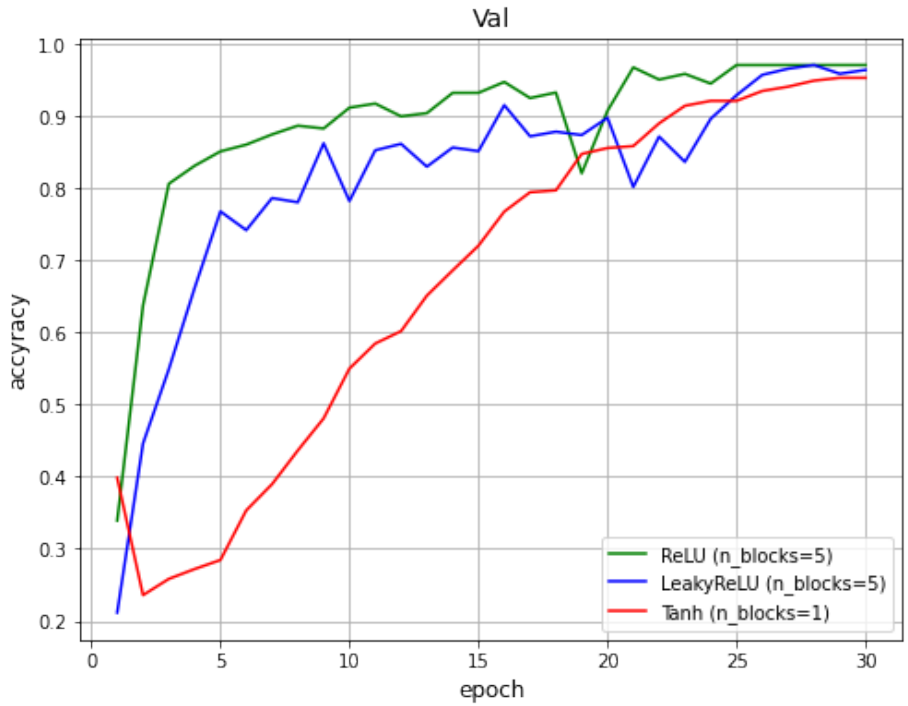
Характеристики компьютера:

PU	name	number of cores
CPU	Intel(R) Xeon(R) CPU 2.30GHz	2
GPU	Tesla T4	2560 (15079MiB)

Кривые обучения



а) accuracy on train



б) accuracy on test

Сравнение с результатами первого этапа

model	parameters	f1_score
SVC	'C': 100, 'kernel': poly	96.1
DecisionTree	'max_depth': 20	95.8
AdaBoost	'n_estimators': 100	39.7
RandomForest	'n_estimators': 50	96.8
GradientBoost	'n_estimators': 300, 'max_depth': 10	96.8
GradientBoost	'n_estimators': 300, 'max_depth': 10	96.8
LogRegression	'C': 1, 'max_iter': 100, 'penalty': 'none', 'solver': lbfgs	26.45
MLP	'max_iter': 1000, 'hidden_layer_sizes': (100, 3)	95.2
FCNN	'block structure': Linear/Bn/ReLU, 'N blocks': 5, 'N parameters': 45701	98.99

Данные

- 2 колонки: месяц; общее количество человек в США, летавших в этот месяц
- с 1949 по 1960 год

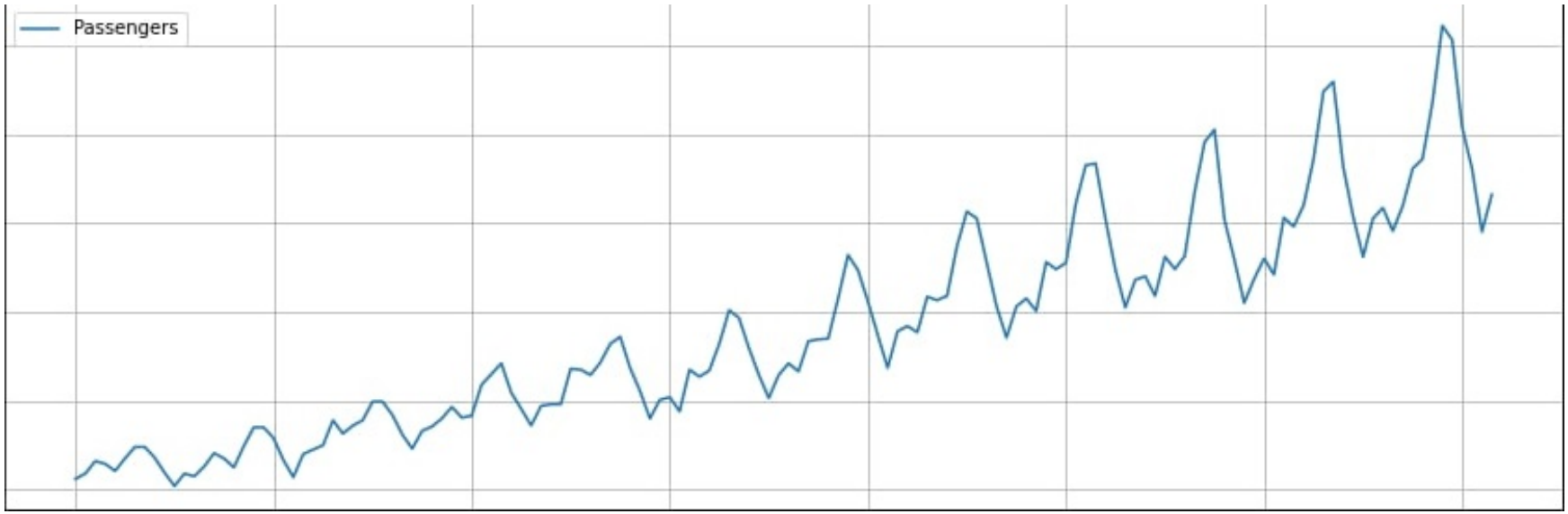


Рис.: Временной ряд

Датасет 2

Параметры, результаты

Результаты для простых моделей:

model	RMSE
mean	122.149
linear	45.7598
seasonal_decompose	9.9855

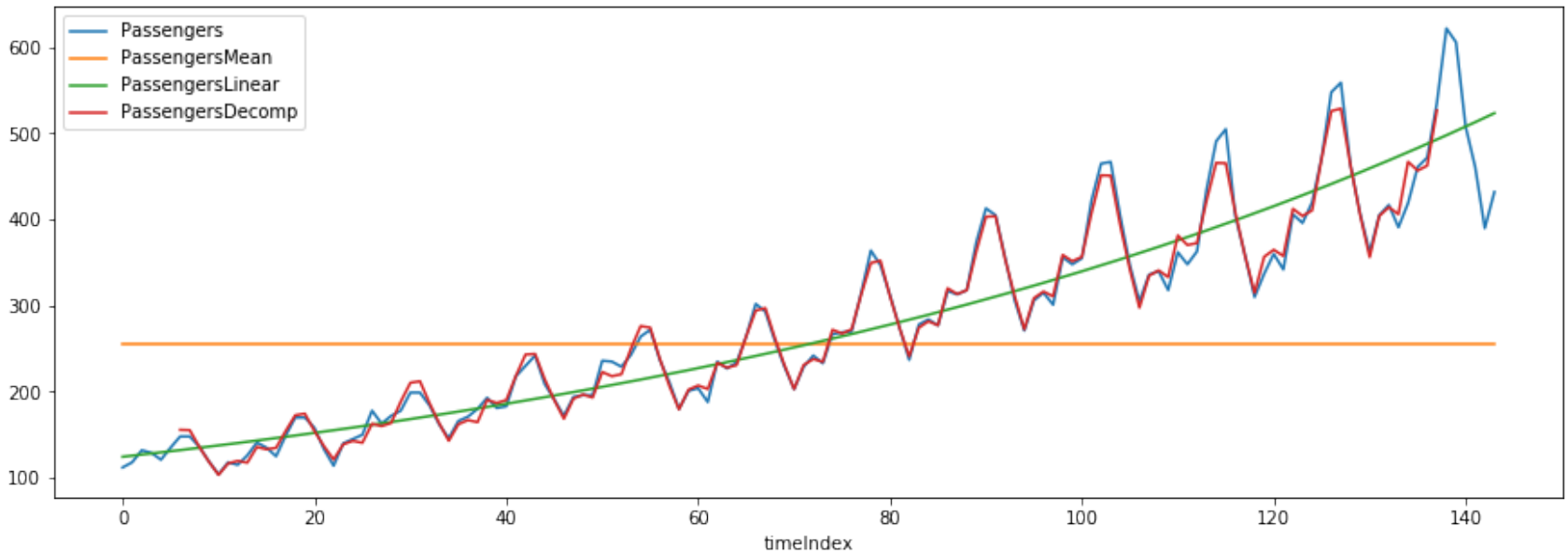


Рис.: Приближение средним, линейное

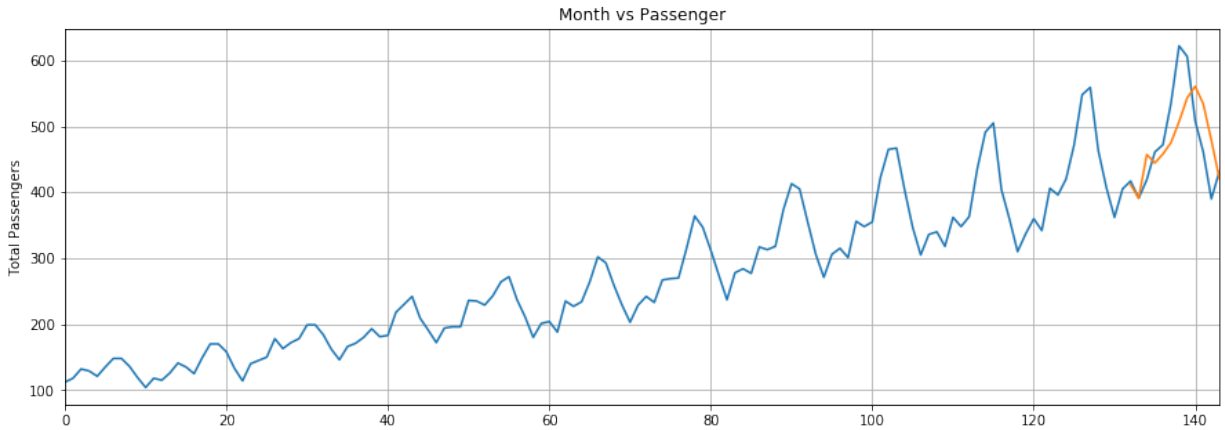
Графики для временного ряда

Параметры LSTM:

- Архитектура: 1 скрытый слой со 100 нейронами
- метод оптимизации Adam
- функция потерь: MSE
- количество эпох: 126
- learning rate: 0.001

Время выполнения для LSTM: train – 44.8 s, test – 0.01 s

Хакактеристики компьютера: RTX 2080Ti, 11264 Mb, Intel Core i9-9900KF, 3.6GHz



Вклад

Канаметов Азамат: использование нейросетевых моделей для задачи классификации.
Пузач Владислав: построение модели логистической регрессии и оформление результатов.
Байшев Олег: прогнозирование временного ряда.

Репозиторий