При работе с данными я использовал библиотеку dask и pandas для загрузки и обьединения данных,для работы с данными я использовал библиотеки sclearn,для взуализации matplotlib.

Обьеденял я по столбцам ‘id ’ и ‘buy_time’ т.к. нужно получить понимание покупки покупателей в определенное время.

Модель для baseline я использовал LogisticRegression, для предсказания на тесте я выбрал catboost т.к. он показал наилучший результат из выбраных мною моделей, таблица с предсказаниями лежит в файле score.csv её при необходимости можно посмотреть отдельно, так же я делал gridsearch на трех моделях catboost, градиентный бустинг и lgboost, но возможно из-за неправильно подобранных параметров результат был не лучше стандартных параметров этих моделей, много эксперементировать не получилось, по причине не большой мощности моего пк, очень долго он обрабатыват сетку параметров, перед предсказанием я делаю стандартизацию т.к. она дала небольшой прирост на окончательной метрике

	f1	LogisticRegression()	KNeighborsClassifier()	GradientBoostingClassifier()	LGBMClassifier()	CatBoostClassifire	RandomForestClassifier()
0	1	0.58	0.68	0.94	0.94	0.94	0.92
1	macro	0.48	0.66	0.94	0.95	0.95	0.93
2	micro	0.48	0.66	0.94	0.95	0.95	0.93
3	weighted	0.47	0.65	0.94	0.95	0.95	0.93

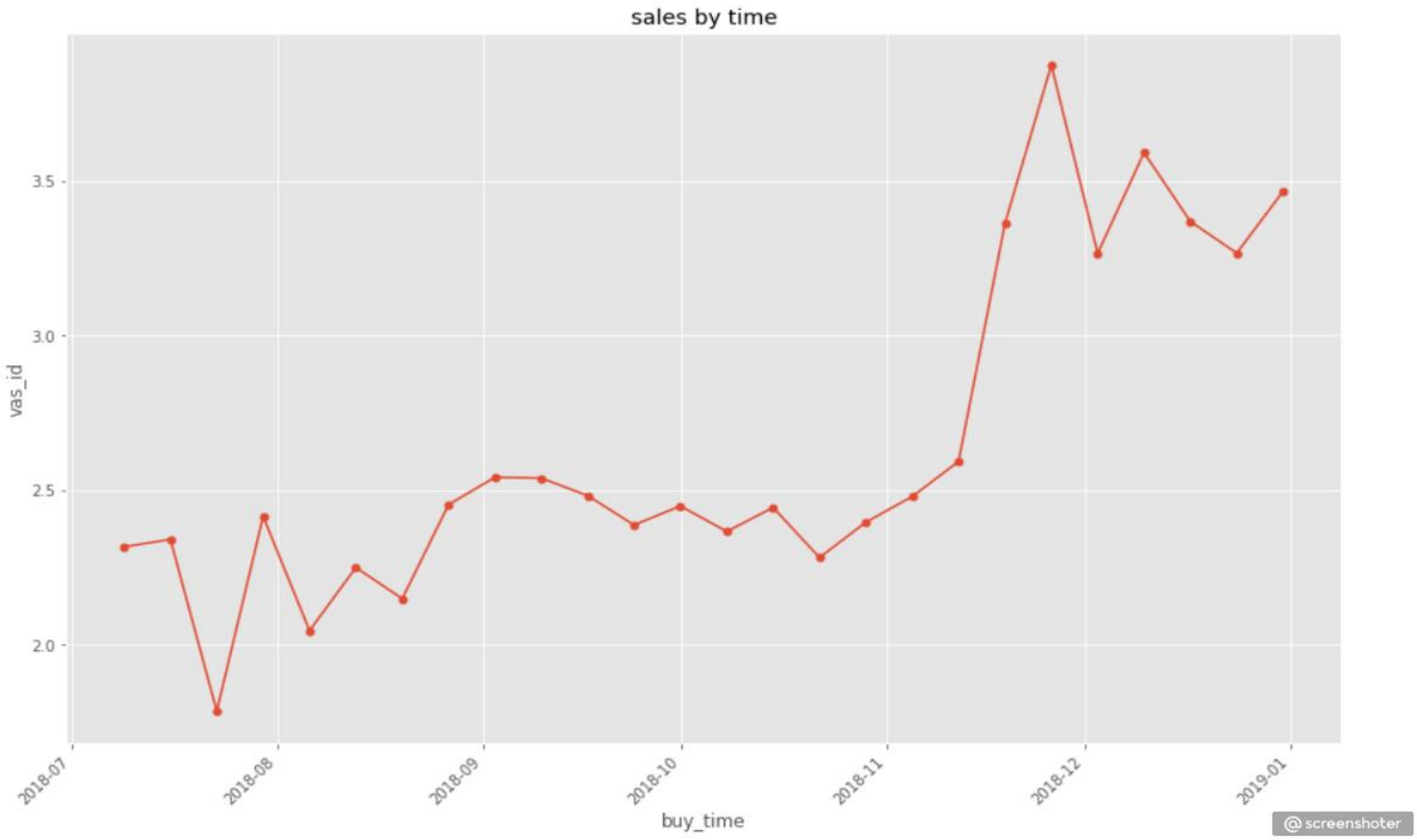

График показывает покупку товаров в определенное время

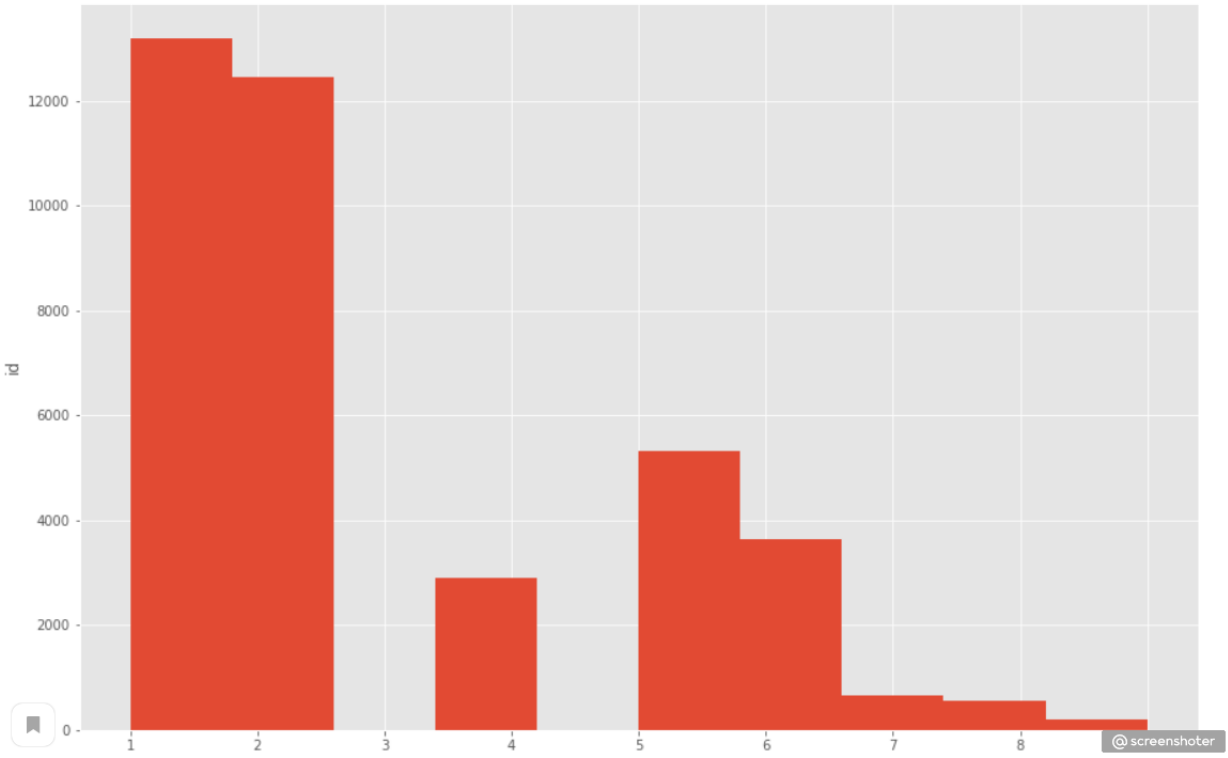

График показывает наиболее популярные продукты у покупателей