Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт математики и информационных технологий

**Отчет по предмету “Технологии обработки данных”**

Выполнила студентка группы 22503

Масаева Ольга

Руководитель курса: Седов

Алексей Владимирович,

кандидат технических наук, доцент

Петрозаводск 2022 г.

*Постановка задачи*: Основным источником задачи является соревнования , производимое на платформе Kaggle [1] “Santander Customer Transaction Prediction”. Цитата из описания соревнования: “Наша миссия в Сантандере — помогать людям и предприятиям процветать. В этом задании мы предлагаем вам помочь нам определить, какие клиенты будут совершать конкретную транзакцию в будущем, независимо от суммы транзакции. Данные, предоставленные для этого конкурса, имеют ту же структуру, что и реальные данные, которые у нас есть для решения этой проблемы.”

Описание данных: Цитата из [1] “ Вам предоставляется анонимный набор данных, содержащий числовые переменные признаков, двоичный целевой столбец и строковый столбец ID\_code. Задача состоит в том, чтобы предсказать значение целевого столбца в наборе тестов.”

*Цель:* Оптимальное решение задачи бинарной классификации данных транзакций.

Для решения данной цели были поставлены следующие задачи:

1. Первичное изучение данных (размеры,статистические характеристики)
2. Графическая иллюстрация характеристик (построение графиков)\
3. Применение корреляционного и факторного анализа признаков
4. Подбор модели и реализация модели для классификации
5. Анализ метрик моделей
6. Формирование выводов

Реализация анализа данных проводилась на языке программирования python, с использованием программной среды Jupiter notebook, а также библиотек numpy, pandas,matplotlib, sklearn, lightgbm, scipy

# Первичное изучение данных

Данные располагались в csv файле train.csv, преобразованы в data frame,отрывок из которого представлен ниже:

ID\_code target var\_0 var\_1 var\_2 var\_3 var\_4 var\_5 var\_6 \

0 train\_0 0 8.255 -6.7863 11.9081 5.0930 11.4607 -9.2834 5.1187

1 train\_1 0 11.5006 -4.1473 13.8588 5.3890 12.3622 7.0433 5.6208

2 train\_2 0 8.6093 -2.7457 12.0805 7.8928 10.5825 -9.0837 6.9427

3 train\_3 0 11.0604 -2.1518 8.9522 7.1957 12.5846 -1.8361 5.8428

4 train\_4 0 9.8369 -1.4834 12.8746 6.6375 12.2772 2.4486 5.9405

var\_7 ... var\_190 var\_191 var\_192 var\_193 var\_194 var\_195 \

0 18.6266 ... 4.4354 3.9642 3.1364 1.6910 18.5227 -2.3978

1 16.5338 ... 7.6421 7.7214 2.5837 10.9516 15.4305 2.0339

2 14.6155 ... 2.9057 9.7905 1.6704 1.6858 21.6042 3.1417

3 14.9250 ... 4.4666 4.7433 0.7178 1.4214 23.0347 -1.2706

4 19.2514 ... -1.4905 9.5214 -0.1508 9.1942 13.2876 -1.5121

var\_196 var\_197 var\_198 var\_199

0 7.8784 8.5635 12.7803 -1.0914

1 8.1267 8.7889 18.3560 1.9518

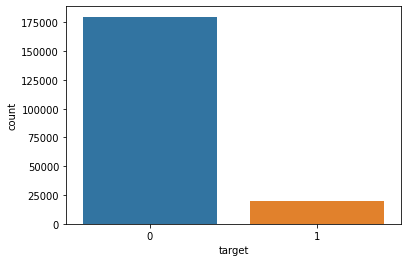
2 -6.5213 8.2675 14.7222 0.3965

3 -2.9275 10.2922 17.9697 -8.9996

4 3.9267 9.5031 17.9974 -8.8104

В обучающей выборке 200 вещественных признаков(var\_0-199) и один целевой бинарный признак(target) - имеется задача бинарной классификации. Всего в выборке 200000 записей.

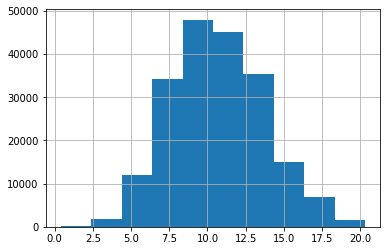
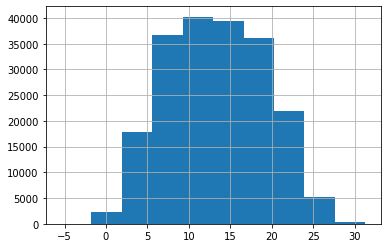
На рисунке 1. представлена гистограмма подсчета целевого признака на имеющихся данных.Видно , что выборка несбалансированная: количество записей со значением 0 примерно в 9 раз больше чем со значением 1. Статистические данные подтверждают это предположение: доля нулей в целевом признаке составляет 0.89951, в то время как доля единиц составляет 0.10049.

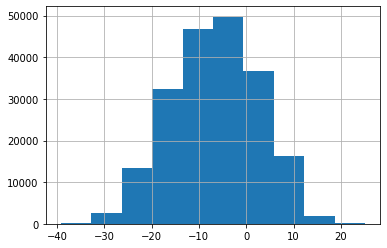
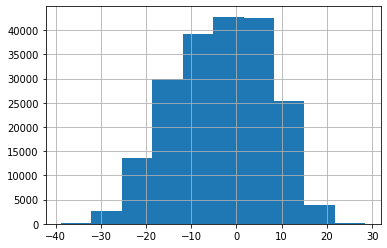
Рис 1. Гистограмма подсчета целевого признака.

На рисунке 2 представлены гистограммы отдельных признаков данных выборки.Визуально графики напоминают нормальное распределение,но при применении к ним тест Шапиро-Вилка на нормальность (функция scipy.stats.shapiro [2]) значения p-value были существенно малы (порядка 1.1724418070891532e-36), что отвергает гипотезу о нормальности распределений.При приближении заметно наличие более чем одного пика у гистограммы, возможно распределение является смесью нормальных распределений.

Также на рисунке 3 приведены гистограммы 1-го признака , разделенные по значению целевой переменной. Также наблюдается визуально сходство гистограмм, но данный факт также требует статистической проверки. Несмотря на внешнюю схожесть распределений, тест на однородность Колмогорова -Смирнова ( функция ks\_2samp [3]) они не прошли , так как значение pvalue ниже 1% .

На рисунке 4 представлены графики распределения математического ожидания и среднеквадратического отклонения признаков.





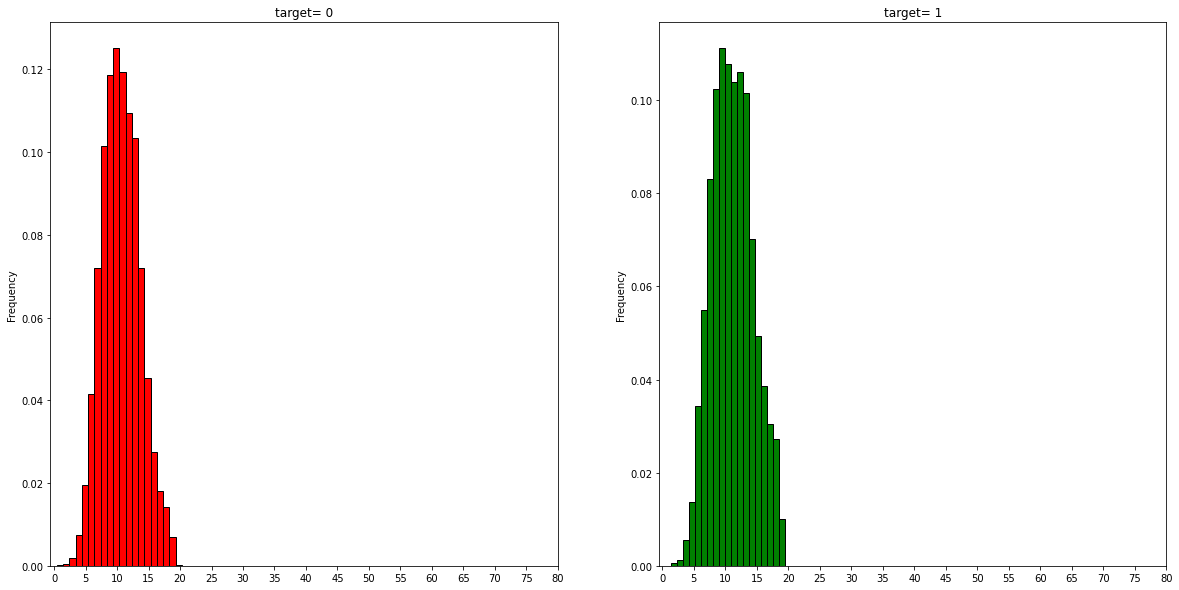
Рис 2.Графики гистограмм распределения параметров 1(левый верхний),20(правый верхний),100(левый нижний),199 ( правый нижний)

Рис 3. гистограммы признака var\_0, разделенного по значению target.

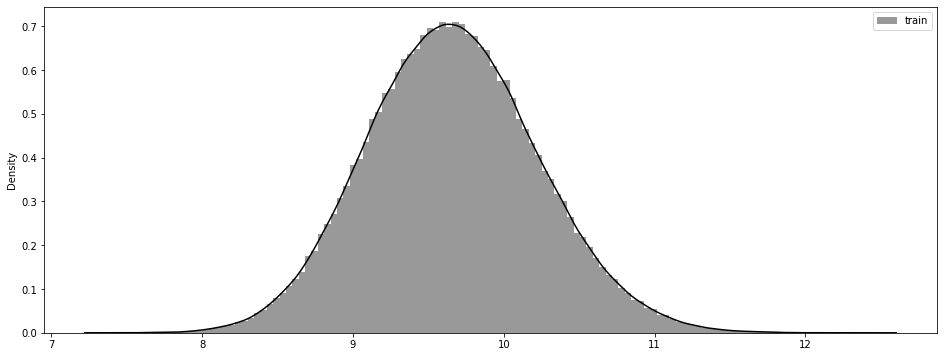
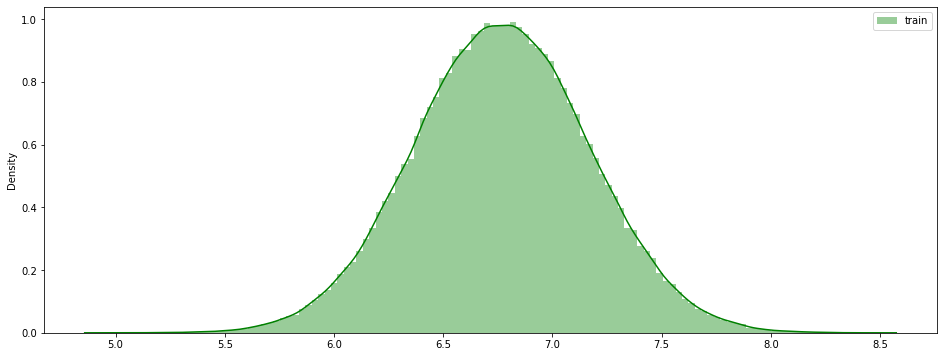


Рис 4. Графики Распределений математических ожиданий и среднеквадратических отклонений признаков

# Корреляционный и Факторный анализ

В качестве корреляционного анализа был подсчитан парный коэффициент корреляции Пирсона. Без учета идентичных пар, наименьшее значение корреляции наблюдается между парой var\_75 и var\_191 cor = 2.703975e-08, наибольшее значение корреляции наблюдается между парой var\_139 и var\_26 cor = 0.009844. Учитывая тот факт , что наибольшая парная корреляция является крайне малой величиной, можно сделать вывод о независимости признаков.

Также, учитывая большой объем данных, рассматривалось использование метода главных компонент(PCA алгоритм [4] со стандартными параметрами и n\_components=2 ) для уменьшения размерности вектора признаков. Результатом сжатия данных к двух компонентам стал вектор дисперсий “Results of variance in 2 columns used : [0.005654 0.00514221]” ,что в сумме составляет 0.010796214605994834. Так как потеря данных незначительная - 0.01 , метод главных компонент имеет место быть Результат уменьшение размерности представлен на рисунке 5. - Фиолетовые точки соответствуют точкам с target = 0, желтые c target = 1



Рис. 5. График скопления точек выборки в новой системе координат

# Использование моделей

В качестве моделей было взято 2 модели - логистическая регрессия и LGBM. Обе модели показали неплохие результаты и работали достаточно быстро даже без сокращения размерности

В соревновании [1] в качестве метрики качества использовалась ROC AUC.

*Логистическая регрессия*:

Логистическая регрессия или логит-модель (англ. logit model) — статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путем его сравнения с логистической кривой. Эта регрессия выдает ответ в виде вероятности бинарного события (1 или 0). Эта модель часто применяется для задач классификации. Для нашей модели данных использовалась функция LogisticRegression ([5]) c дефолтными параметрами, в качестве данных выступал полный набор датасета. Результаты метрик на валидационной выборки представлены ниже:

Confusion\_matrix:

[[35625 435]

[ 3033 907]]

Accuracy: 0.9133

Precision: 0.6759

Recall: 0.2302

F1: 0.3535

AUC: 0.6091

Алгоритм дал большую долю правильных ответов, поэтому имеет высокую рассчитанную точность, но , так как размеры классов не сбалансированные , по данной метрике не стоит судить о работе алгоритма. Точность имеет среднее значение,а полнота и вовсе низкое, за счет чего F метрика также имеет низкое значение. Значение AUC примерно 0.6, а значит качество алгоритма лучше случайного выбора класса,но не намного

Для сжатой методом главных компонент модели логистической регрессии результаты метрик представлены ниже:

Confusion\_matrix:

[[36013 40]

[ 3888 59]]

Accuracy: 0.9018

Precision: 0.5960

Recall: 0.0149

F1: 0.0292

AUC: 0.5069

Видно, что во второй класс в этой модели классифицировалось куда меньшее число элементов ,чем в предыдущей модели. Значения характеристик уменьшились, особенно полнота. Значение площадь под кривой ROC, говорит о том , что данный выбор не сильно лучше случайного выбора.

*Light Gradient Boosted Machine*

В машинном обучении классификатор LightGBM является частью семейства Boosting, и сегодня это самая распространенная модель классификации в сообществе машинного обучения. LightGBM – это мощная модель машинного обучения, которую можно формировать в зависимости от задачи, над которой вы работаете.

Light GBM - это быстрый , распределенный, высокопроизводительный алгоритм градиентного форсирования, поддерживаемый деревом решений, используемый для ранжирования, классификации и других задач машинного обучения.

Для текущей задачи использовался полный набор данных и модель lightgbm[6] с параметрами:

params = {

'device' : 'cpu',

'n\_estimators': 7000,

'num\_leaves': 20,

'max\_depth': -1,

'min\_data\_in\_leaf': 80,

'learning\_rate': 0.008,

'boosting': 'gbdt',

'objective': 'binary',

'metric': 'auc',

'n\_jobs': -1

}

Время обучения составило несколько минут. На валидационной выборке результаты метрик оказались таковыми:

Confusion\_matrix:

[[17779 133]

[ 1463 625]]

Accuracy: 0.9202

Precision: 0.8245

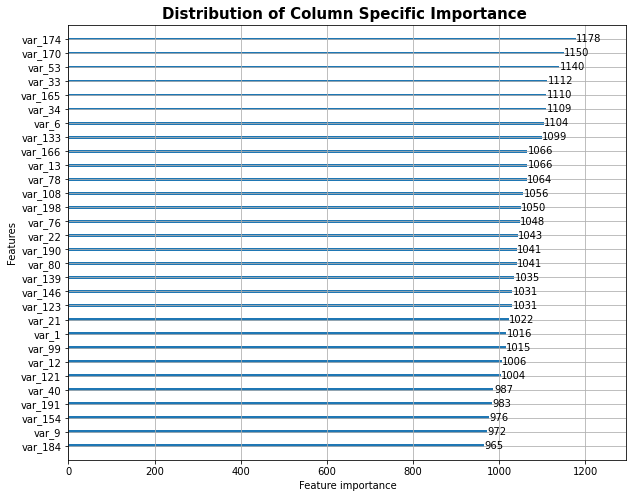
Recall: 0.2993

F1: 0.4392

AUC: 0.6460

Результаты метрик лучше , чем у модели логистической регрессии. Высокая точность, при небольшой полноте дает среднюю оценку F-метрики. Результаты модели на скрытых тестовых данных при посылке в соревнование [1] - 0.89 , весьма неплохой результат.

Также в качестве демонстрации модели можно привести график на рис 6.

рис. 6 график иллюстрирующий важность каждой переменной в построенной модели

# Вывод

В результате работы был произведен анализ данных транзакций и на его основе выбрана модель бинарной классификации. Также было рассмотрено влияние сжатие данных на качество работы модели на примере сжатия методом главных компонент и модели логистической регрессии

Используемые источники:

1. Интернер -ресурс [ссылка: <https://www.kaggle.com/competitions/santander-customer-transaction-prediction> ]
2. Интернет-ресурс [ссылка: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html> ]
3. Интерне-ресурс [ ссылка: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html> ]
4. интернет-ресурс [ссылка: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html> ]
5. интернет-ресурс [ ссылка: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html> ]
6. интернет-ресурс[ссылка: <https://lightgbm.readthedocs.io/en/latest/Python-API.html>]