Прогнозирование оттока клиентов

Цели и задачи проекта:

Отток пользователей - одна из наиболее актуальных задач в областях, где распространение услуги составляет порядка 100%. Ярким примером такой области является сфера телекома. Поэтому особенно важным становится именно удержание клиентов, а не привлечение новых. Кроме того, поскольку данные пользователей обфусцированы, мы не имеем представления о значении и роли каждой переменной в данных.

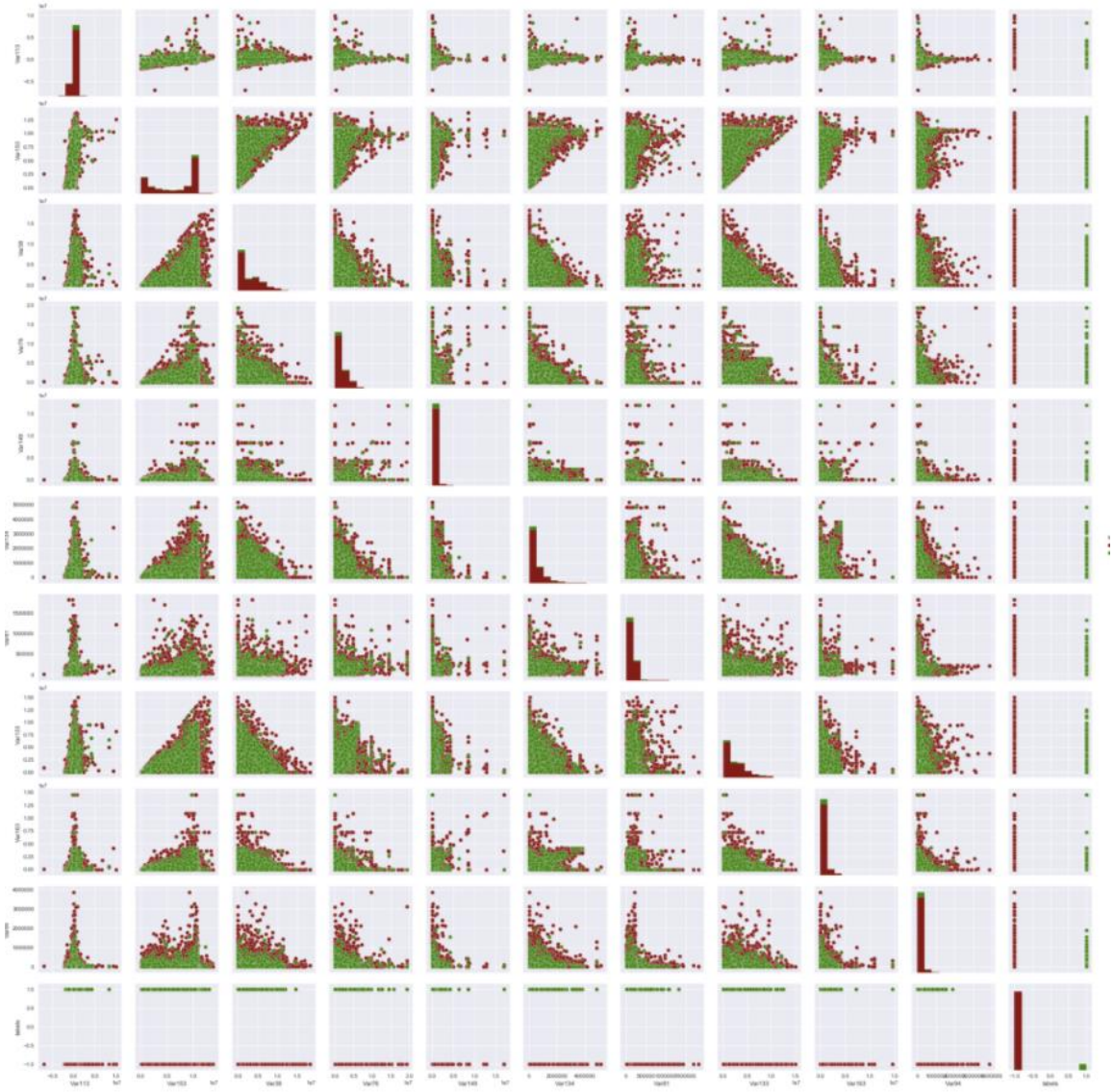
В терминах машинного обучения мы решаем задачу бинарной классификации: отток и не отток.

Полученная система позволит находить клиентов, склонных к оттоку (телеком-оператора) по их поведению и последующих мероприятий, связанных с удержанием клиентов. Примером таких мероприятий может быть обзвон клиентов и предложение более выгодного тарифного плана или скидка на текущий тарифный план.

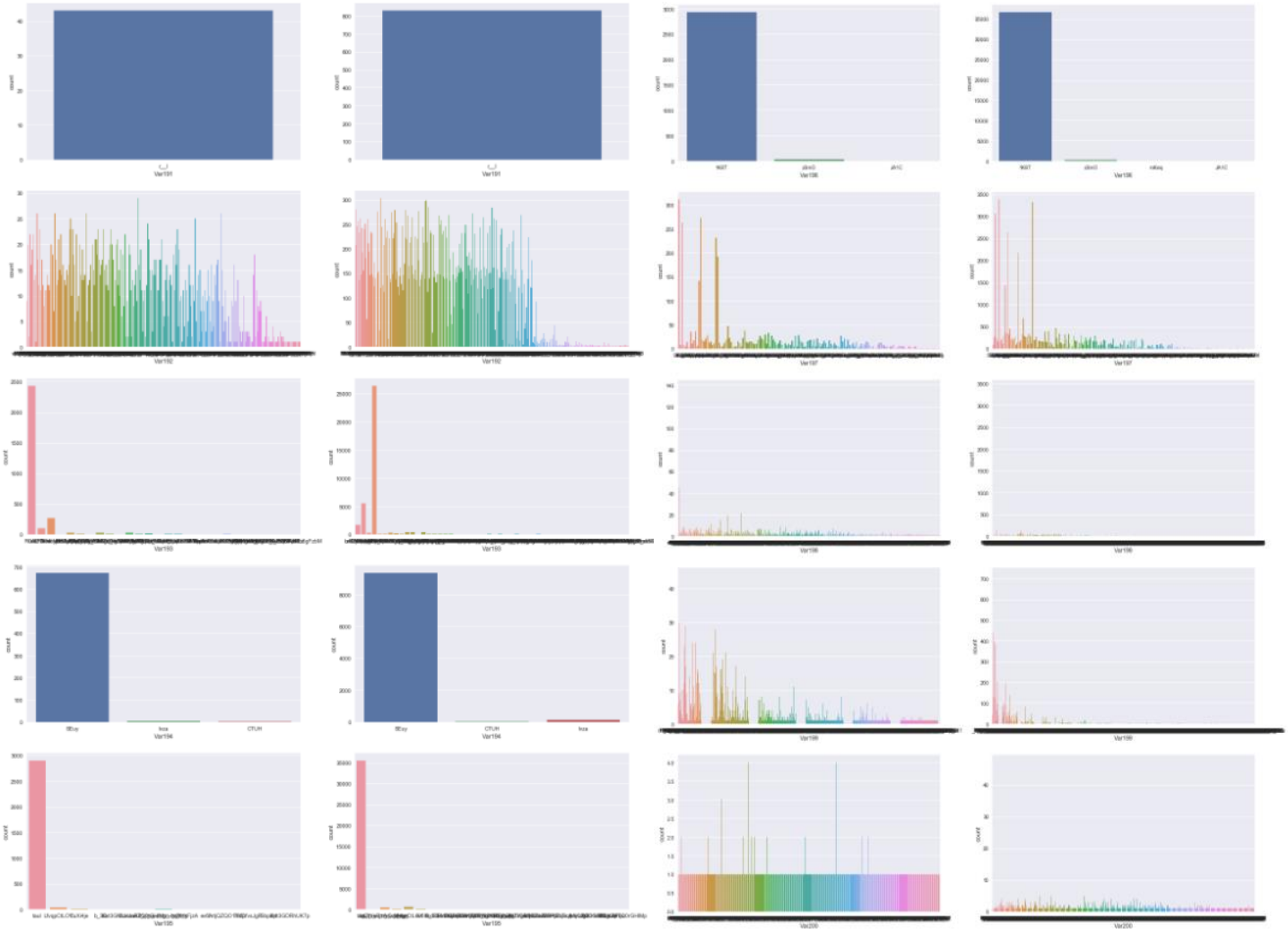
Описательный анализ данных:

На этапе описательного анализа данных был выявлен большой дисбаланс по классам:
отток: 7.44%
не отток: 92.56%

Распределение в разрезах классов наиболее коррелирующих непрерывных переменных показало большое количество выбросов и малое количество точек из-за сильной разреженности данных:



Гистограммы категориальных признаков в разрезе классов показали огромное количество категорий, что затрудняет их обработку:



Методика измерения качества и критерий успеха:

Из-за несбалансированности классов имеет смысл использовать метрику ROC-AUC. Метрики такие как PR-AUC, accuracy, precision, recall смысла использовать нет, так как эти метрики очень чувствительны к несбалансированности классов. Процесс тестирования модели лучше проводить на новых данных, возможно AB-тестирование. Тестировать модель лучше всего рассматривая экономический эффект, так как экономический эффект от внедрения модели будет показывать реальный результат работы, так же это позволит получить детальную информацию о цене ошибок первого и второго рода. Критерием успешности можно считать увеличение площади ROC-кривой на 15-20% и увеличение экономического эффекта после AB-тестирования.

Техническое описание решения:

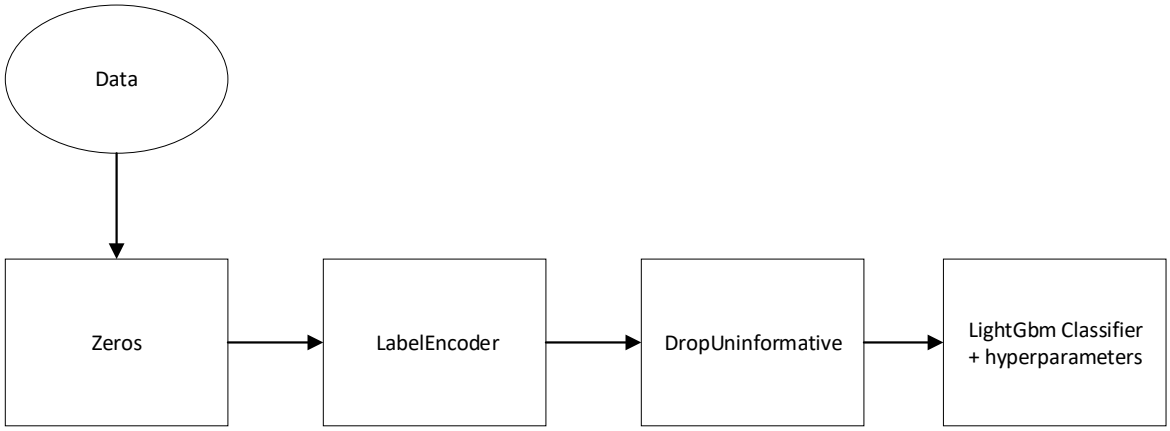
В ходе эксперимента были опробованы следующие стратегии:

1. Балансировка выборки: **underdampling**, **oversampling**.
2. Заполнение пропущенных значений: **mean**, **zeros, median**, **most frequent**.
3. Обработка категориальных признаков: **Label encoding**, **One-hot encoding**, **Frequency encoding**.
4. Регуляризация: **Lasso**, **Ridge**.
5. Классификаторы: **RidgeClassifier**,**LoghisticRegression, RandomForestClassifier**, **GradientBoostingClassifier**. Наилучшее качество показал **GradientBoosting**.
6. Настройка с помощью **GridSearchCV** гиперпараметров **GradientBoosting** из библиотеки **LightGBM**.

На основе экспериментов ниже представленная цепочка обработки данных показала наилучшее качество, так что она была выбрана для построения модели.

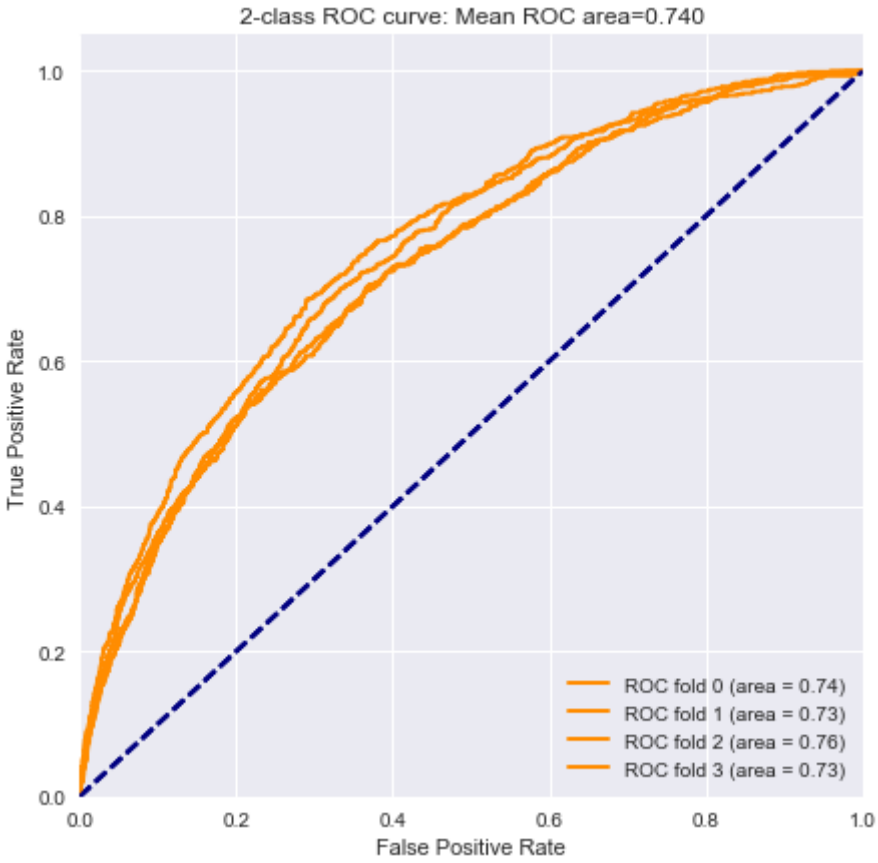
Pipeline:

1. Заполняем пропуски в данных нулями.
2. Обрабатываем категориальные признаки с помощью LabelEncoder.
3. Удаляем неинформативные фичи, в которых одно константное значение.
4. Обучаем LightGBM Classifier с найденными оптимальными параметрами на обработанных данных.



Оценка качества модели:

Модель будет оцениваться по AUC-ROC, итоговая оценка будет проводиться на отложенной выборке, кросс-валидация на обучении – 4 фолда:



На отложенной выборке AUC score: 0.738164.
AUC-ROC score на отложенной выборке не сильно упал, следовательно можно судить о хорошо построенной модели.
Следующие переменные внесли наибольший вклад в модель:

	importance	labels
114	96	Var125
101	35	Var112
199	32	Var216
172	29	Var188
63	23	Var73
192	18	Var209

После построения модели для расчета экономического эффекта можно будет провести следующий эксперимент:

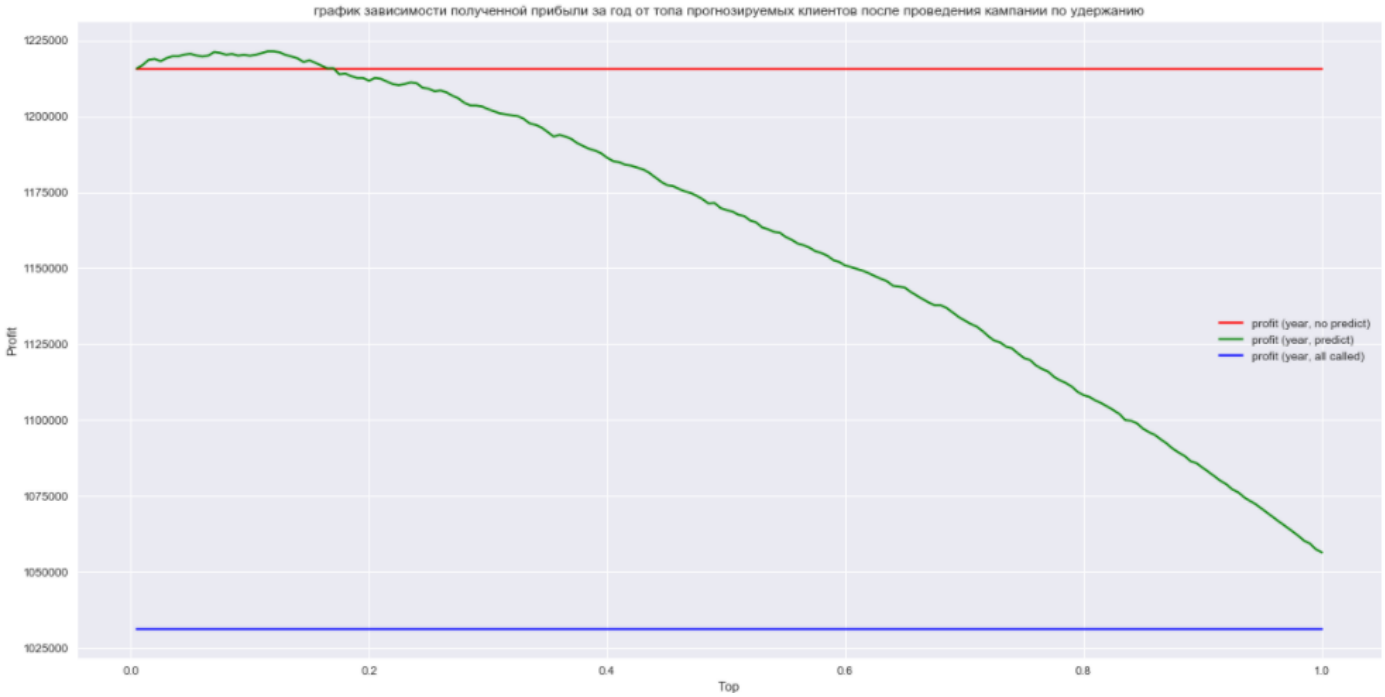
Возьмем 3 группы клиентов:

1. Для которых мы будем применять предиктивную модель и проводить мероприятия по удержанию.
2. Для которых мы не будем делать ничего.
3. Для всех клиентов мы будем проводить мероприятия по удержанию.

Исходные данные эксперимента:

1. Процент конверсии пользователей - 20%;
2. цена тарифного плана после окончания скидки 31р;
3. расходы предприятия на контакт с пользователем + затраты на работу сотрудников 1.3р на одного абонента;
4. предлагаем 20% скидку пользователям, склонных к оттоку;
5. процент согласившихся пользователей, после мероприятий по удержанию 60%.
6. известно, что компания должна нарастить базу на 10% от начального количества абонентов. Для этого компании нужно добрать абонентов до прошлогоднего уровня после кампании по удержанию и добавить 10%
7. шанс, с которым новый абонент согласится на услугу - 50%
8. Затраты на нового абонента - 1.3р
9. всем новым абонентам предлагаем скидку в 50%(15,5р)
10. изначально в эксперименте участвуют 3 группы по 3770 клиентов.

- для нужного нам процента конверсии обрезаем тестовую выборку в соотношении 1:4(отток/не отток)
- расходы считались как сумма скидки для удерживаемого пользователя + затрата на звонок
- в среднем у работника уходит 15 минут на разговор с клиентом, зарплата у сотрудника 1000р. т.е. затрата на звонок = зарплата/рабочие*дни/рабочие*часы/минуты*время_звонка -> 1.3р
- давать прошлогоднюю скидку старому клиенту невыгодно, поэтому даем меньшую (20%)
- для подсчета необходимо просчитать прибыль за первый месяц после проведения кампании по удержанию. Далее мы просчитываем прибыль за следующие месяцы. За каждый месяц равными долями проводится кампания по привлечению новых клиентов.



- годовая прибыль после проведения единоразовой кампании по удержанию клиентов: 1215674.2
- годовая прибыль без проведения единоразовой кампании по удержанию клиентов: 1221443.2
- годовая прибыль после предложения удержания всех клиентов: 1031195.53333

Выводы:

Для данного эксперимента размер оптимального топа составляет 10-12%, при таком размере экономический эффект максимальный.

Из проведенного эксперимента можно сделать вывод, что внедренная модель показывает хороший экономический эффект: при выборе топа 10-12% мы получаем прибыль примерно в 4000р относительно группы клиентов, над которыми кампания не проводилась. Минимальную прибыль показывает группа, над которой проводилась полная кампания по удержанию.