Прогноз вероятности пролонгации полиса

Русланцев А.Н.

Исходные данные

- Исходный массив данных — 96605 клиентов, из них 77407 — обучающая выборка.
- Категориальные признаки:
  - 2 категории: DATA_TYPE, CLIENT_HAS_DAGO, CLIENT_HAS_OSAGO, POLICY_COURT_SIGN, POLICY_HAS_COMPLAINTS, POLICY_HAS_COMPLAINTS, VEHICLE_IN_CREDIT, INSURER_GENDER, POLICY_IS_RENEWED
  - Несколько категорий: VEHICLE_MAKE, VEHICLE_MODEL, POLICY_BEGIN_MONTH, POLICY_END_MONTH, POLICY_SALES_CHANNEL, POLICY_SALES_CHANNEL_GROUP, POLICY_INTERMEDIARY, POLICY_BRANCH, POLICY_CLM_N, POLICY_CLM_GLT_N, POLICY_PRV_CLM_N, POLICY_PRV_CLM_GLT_N, CLIENT_REGISTRATION_REGION
- Числовые признаки
  - POLICY_MIN_AGE, POLICY_MIN_DRIVING_EXPERIENCE, POLICY_PRICE_CHANGE, POLICY_YEARS_RENEWED_N, VEHICLE_ENGINE_POWER, VEHICLE_SUM_INSURED, CLAIM_AVG_ACC_ST_PRD, POLICY_DEDUCT_VALUE

Обработка данных

Необходимо исправить следующие неточности в данных:

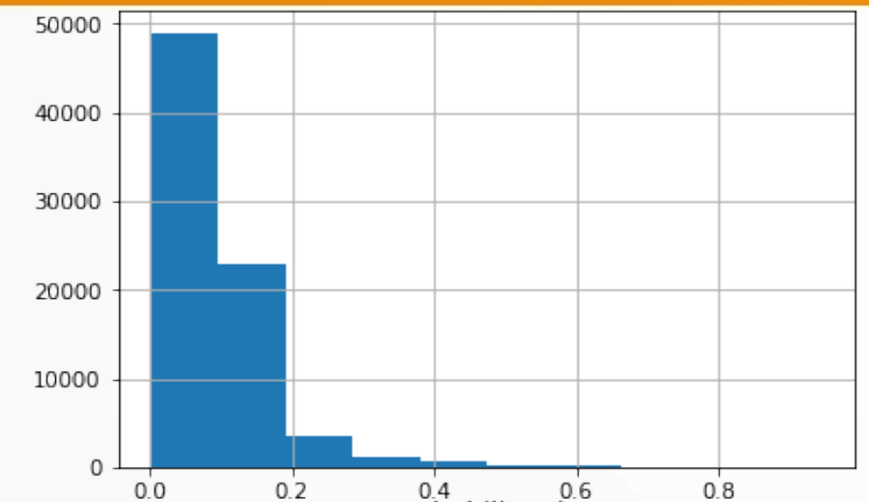
- Производители ВАЗ и Lada — одно и то же
- В некоторых записях вместо опыта вождения стоит год начала стажа
- В данных есть прицепы с ненулевой мощностью двигателя и автомобили с нулевой мощностью
- В данных есть автомобили с мощностью более 1000 л.с.
- Мощность двигателя, сумму франшизы, сумму страхования можно округлить до целых

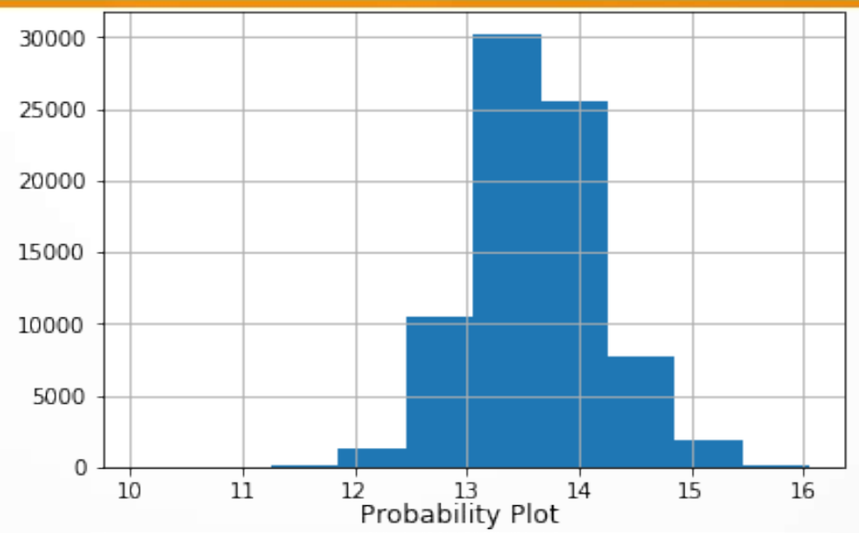
Создание новых признаков

Из исходных данных можно вычислить дополнительно:

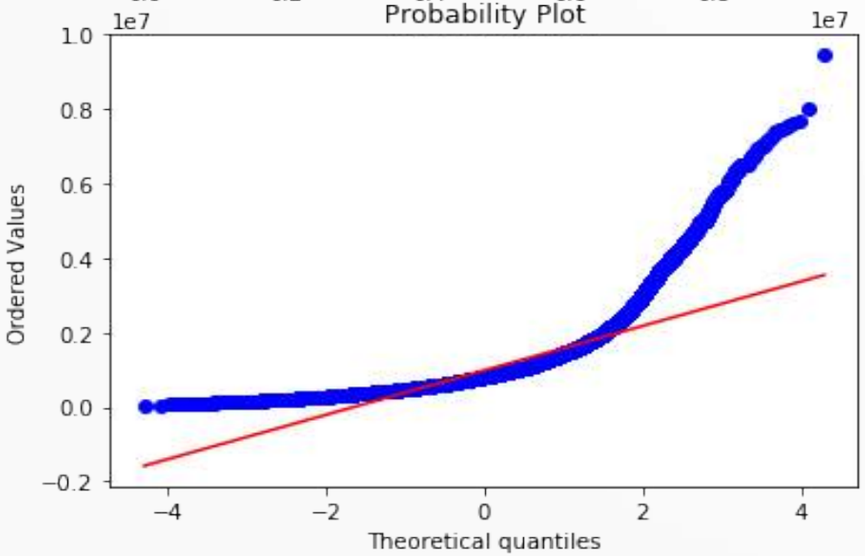
- Была ли куплена страховка на полный год
- Возраст, когда человек получил водительское удостоверение
- Отношение суммы франшизы к премии
- Отклонение франшизы, премии, возраста от среднего по выборке
- Тенденция по участию в ДТП
- Разделить автомобили по мощности, а водителей по стажу и возрасту на категории
- И.т.д.

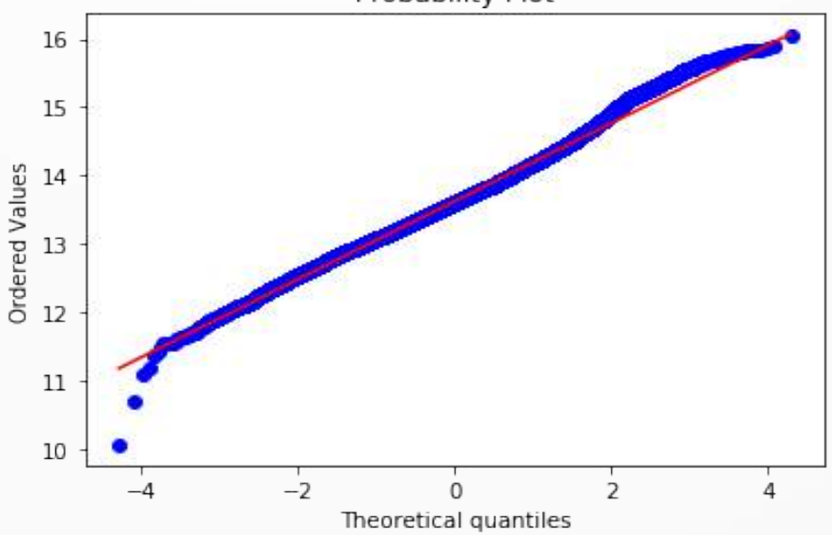
Соответствие числовых признаков нормальному распределению





Некоторые числовые признаки необходимо прологарифмировать, чтобы их распределение было более близким к нормальному





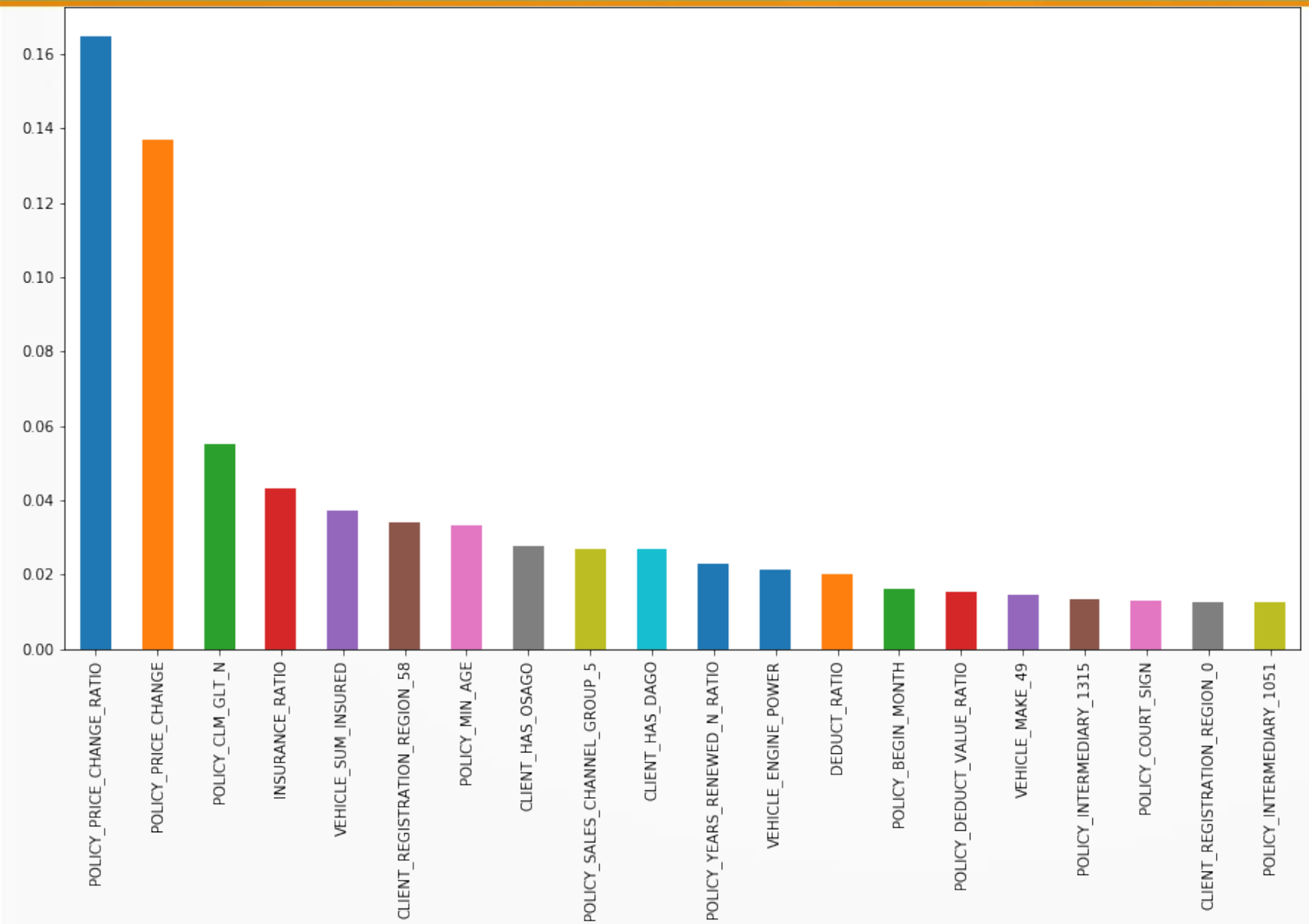
Соответствие признаков нормальному распределению до (слева) и после (справа) логарифмирования

Используемая модель

Были рассмотрены логистическая регрессия, деревья, SVM и градиентный бустинг. Наилучший результат по кросс-валидации показал градиентный бустинг.

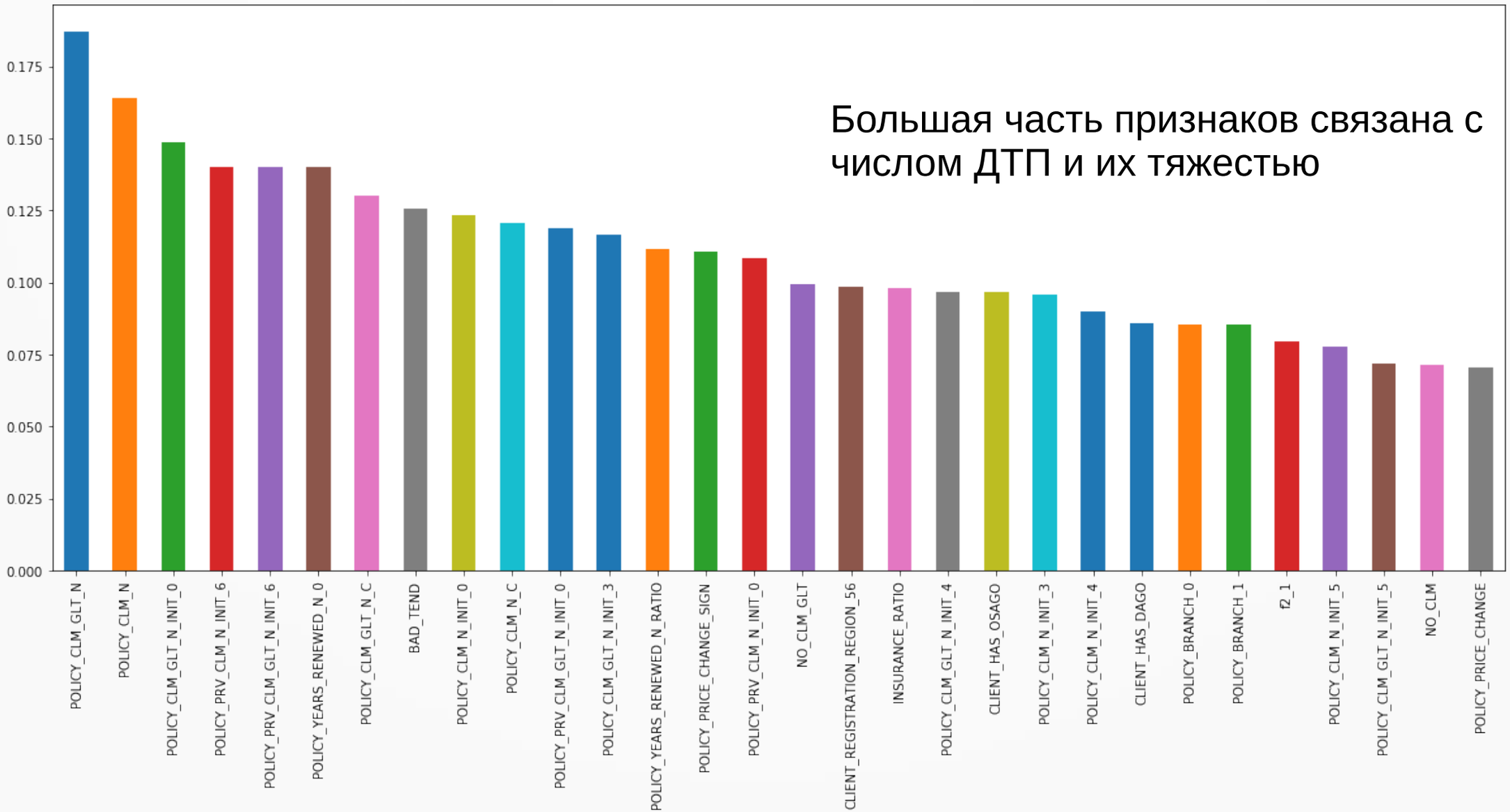
Получено, что из тестовой выборки 15150 клиентов продлят полис, а 4048 — не продлят

20 наиболее важных признаков с точки зрения модели



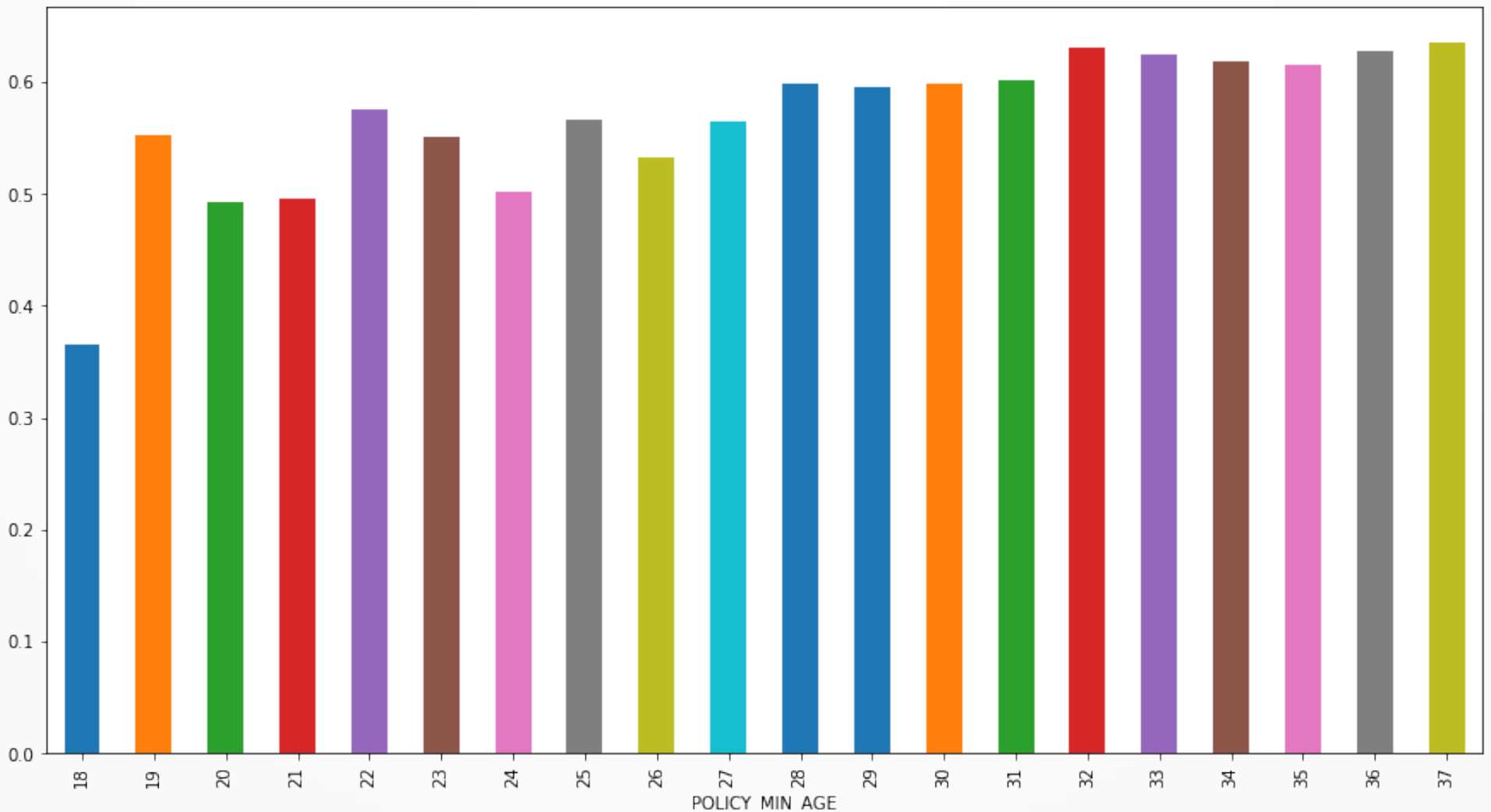
Наиболее важными с точки зрения модели являются изменение цены полиса, знак изменения цены полиса, отсутствие ДТП, длительность страхования в одной компании, стаж, премия, наличие ОСАГО, а так же наличие суда по полису.

Корреляция признаков с целевой переменной



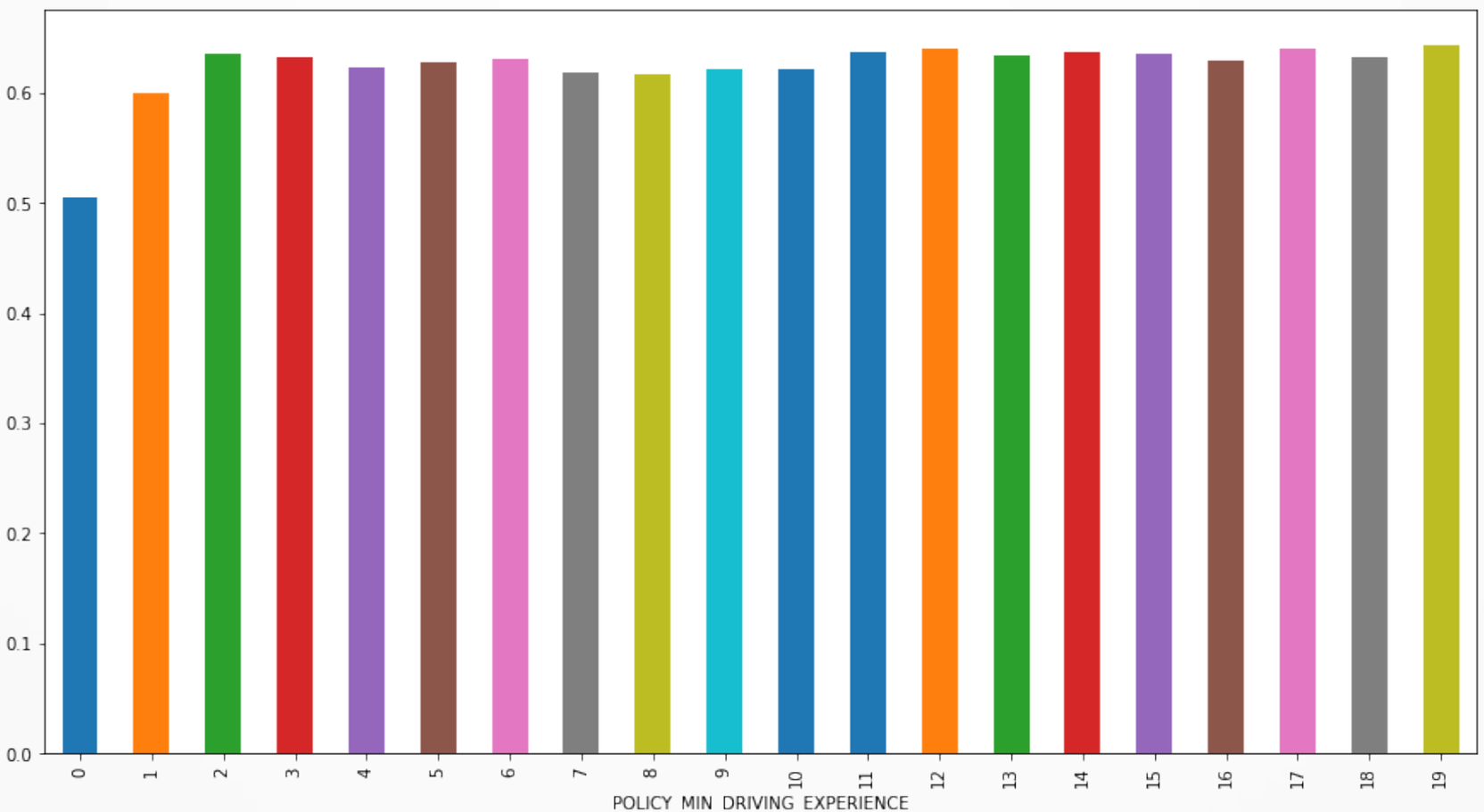
Большая часть признаков связана с числом ДТП и их тяжестью

Вероятность продления полиса в зависимости от возраста водителя



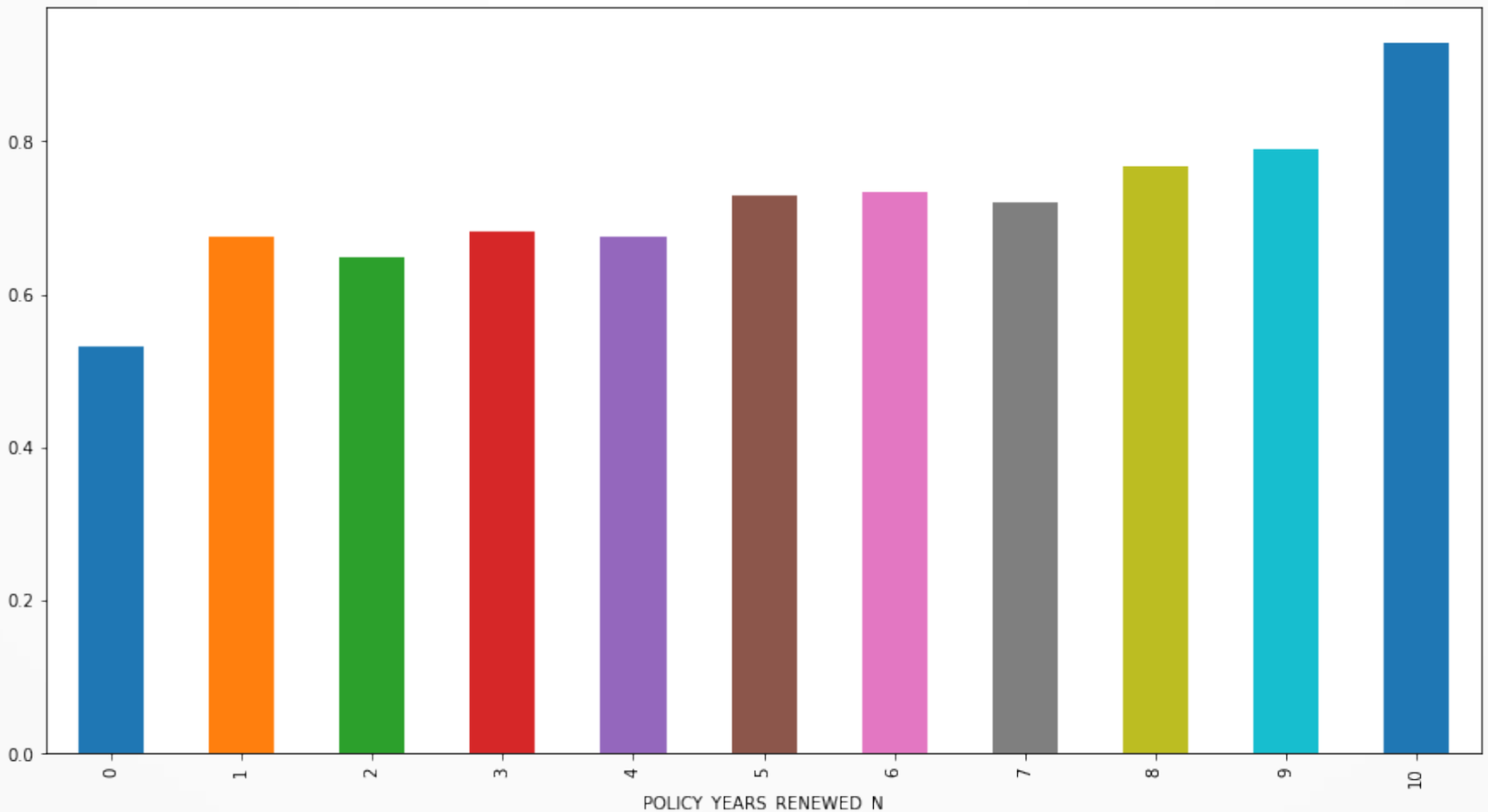
Молодые водители менее охотно продлевают полис

Зависимость вероятности продления полиса от стажа



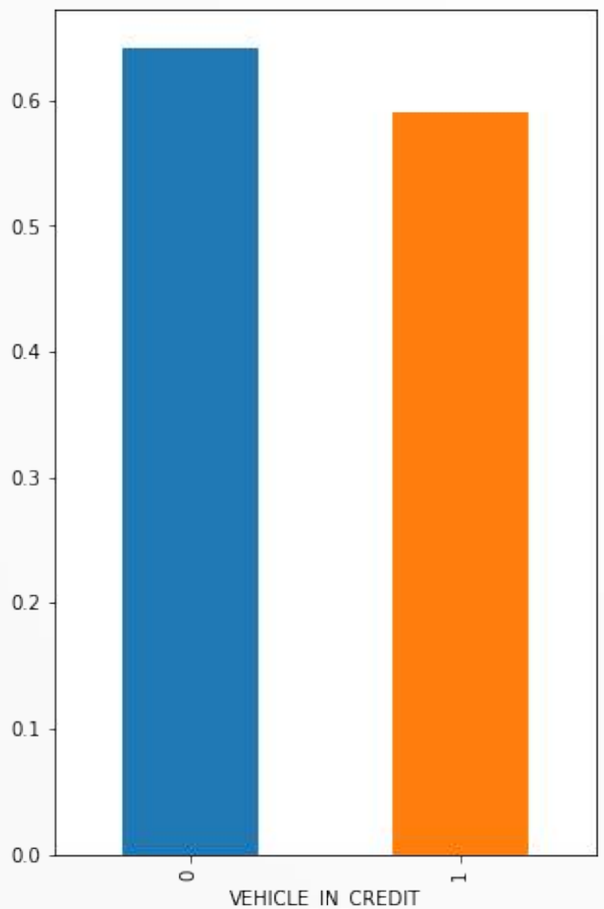
Водители, имеющие стаж от двух лет более охотно продлевают полис

Зависимость продления полиса от количества пролонгаций

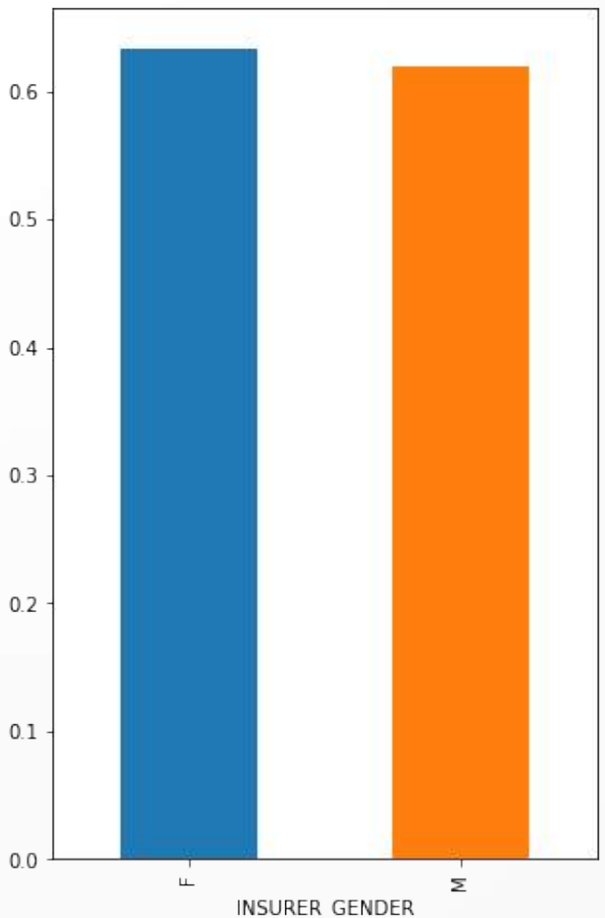


Клиенты, пролонгировавшие полис хотя бы один раз, чаще снова страхуют свой автомобиль

Вероятность продления полиса

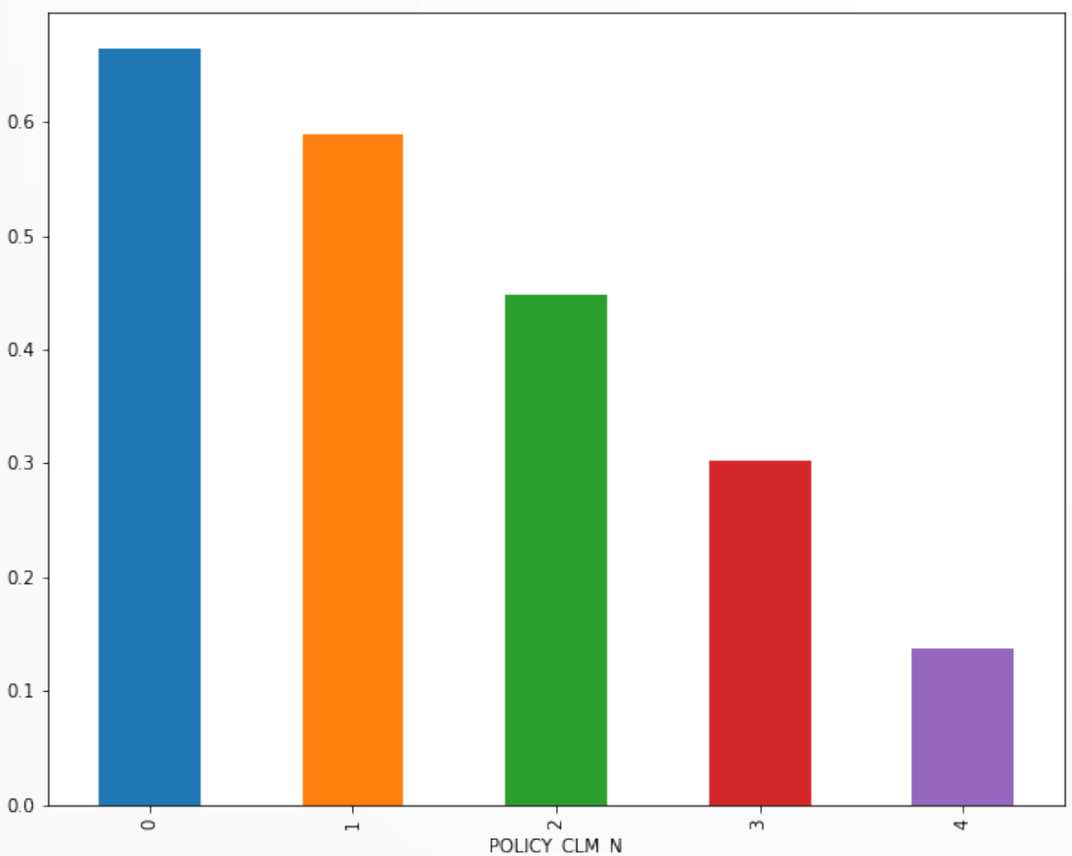


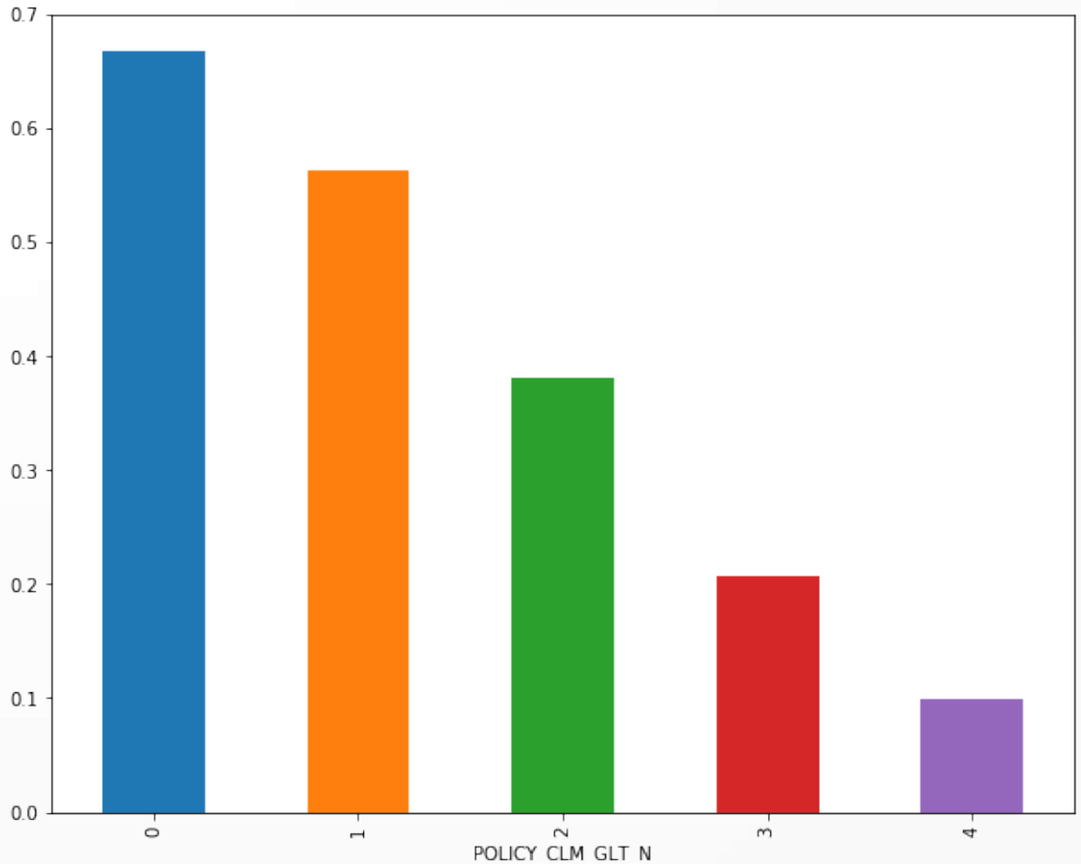
Владельцы кредитных автомобилей с меньшей вероятностью продлевают полис



Мужчины и женщины продлевают полис с примерно одинаковой вероятностью

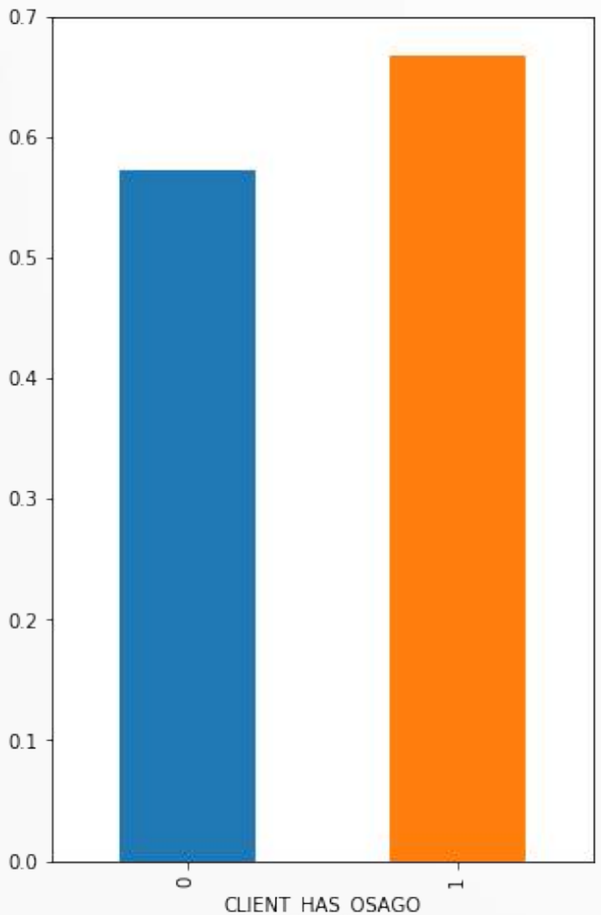
Вероятность продления полиса

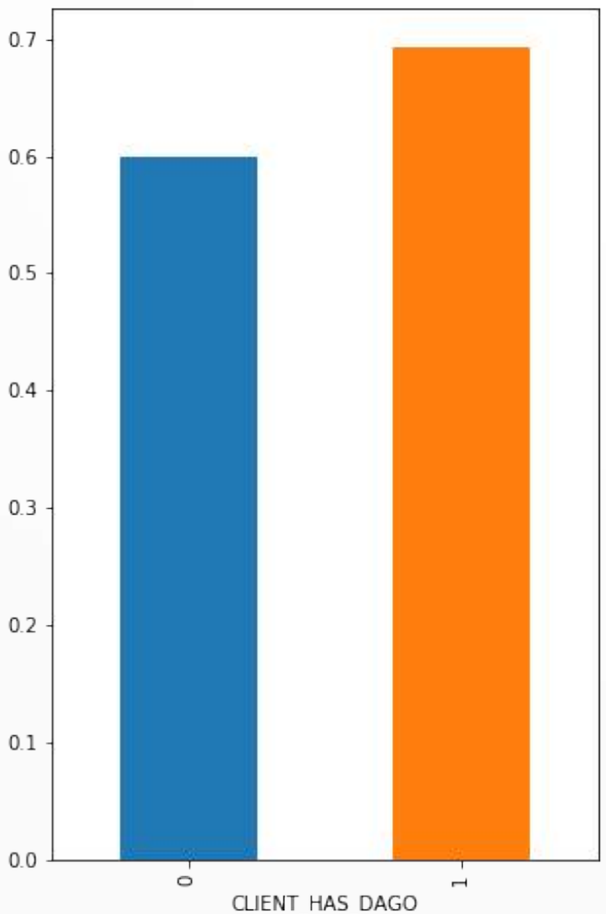




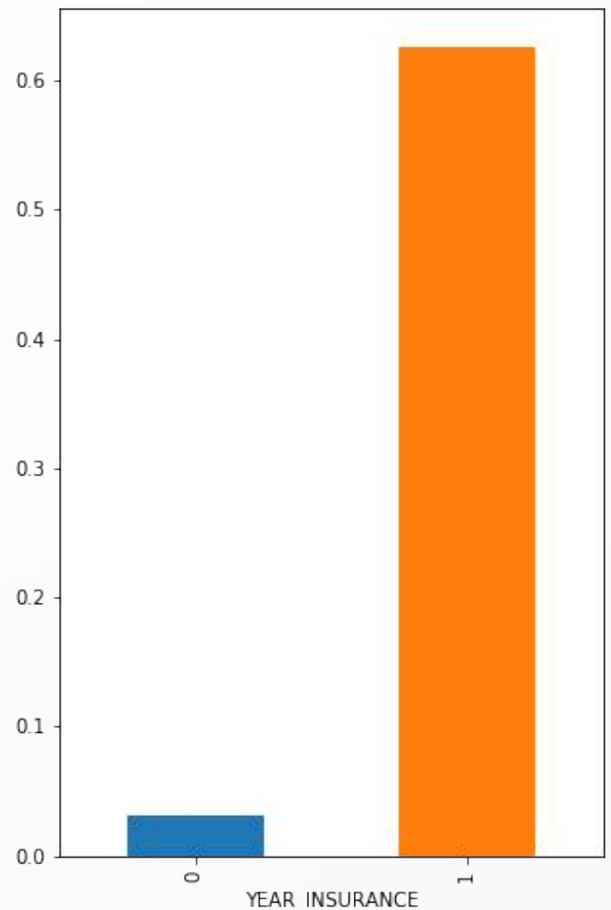
Чем больше число ДТП, тем меньше вероятность продления полиса

Вероятность продления полиса



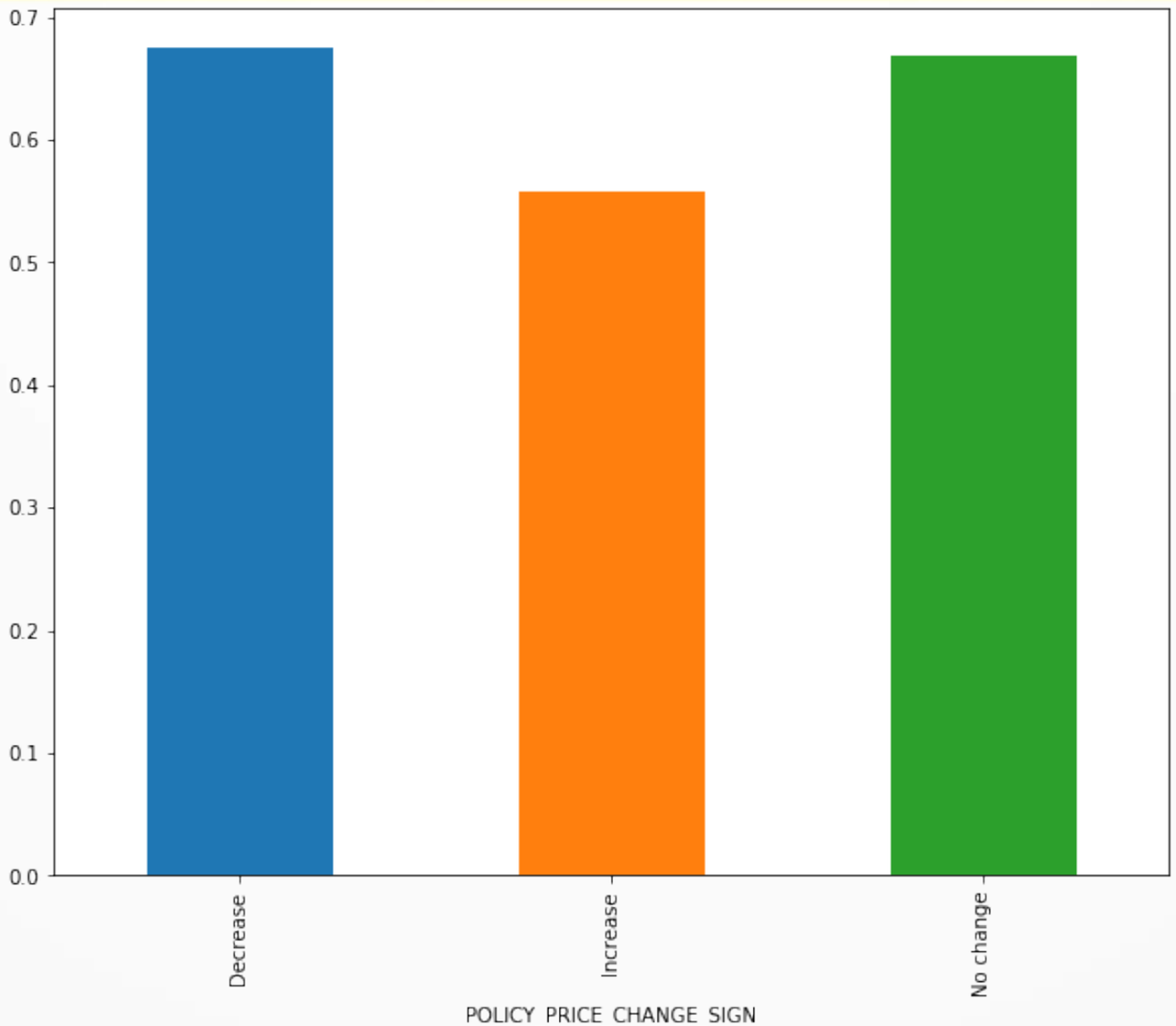


Наличие ОСАГО и ДАГО увеличивает вероятность продления



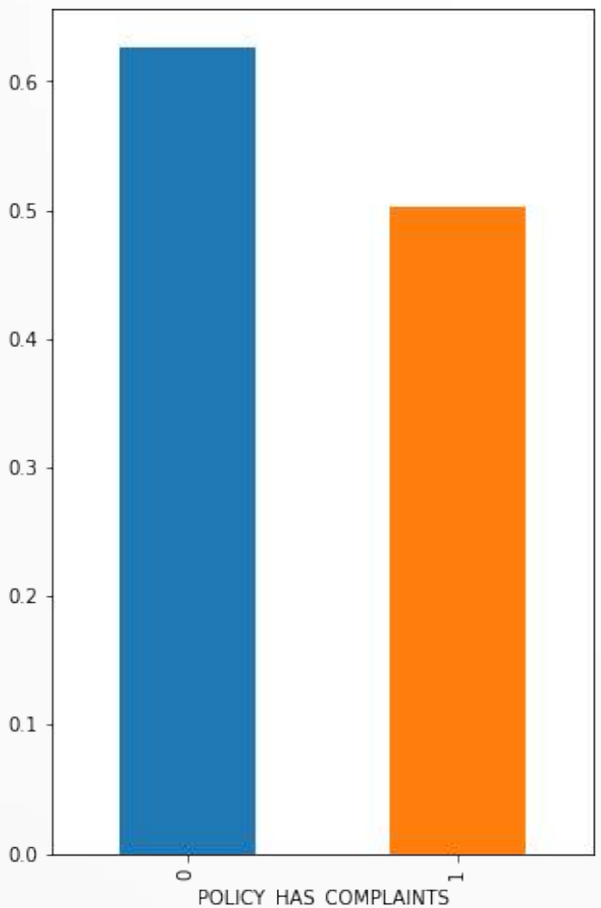
Клиент, купивший полис не на год, вряд ли будет его продлевать

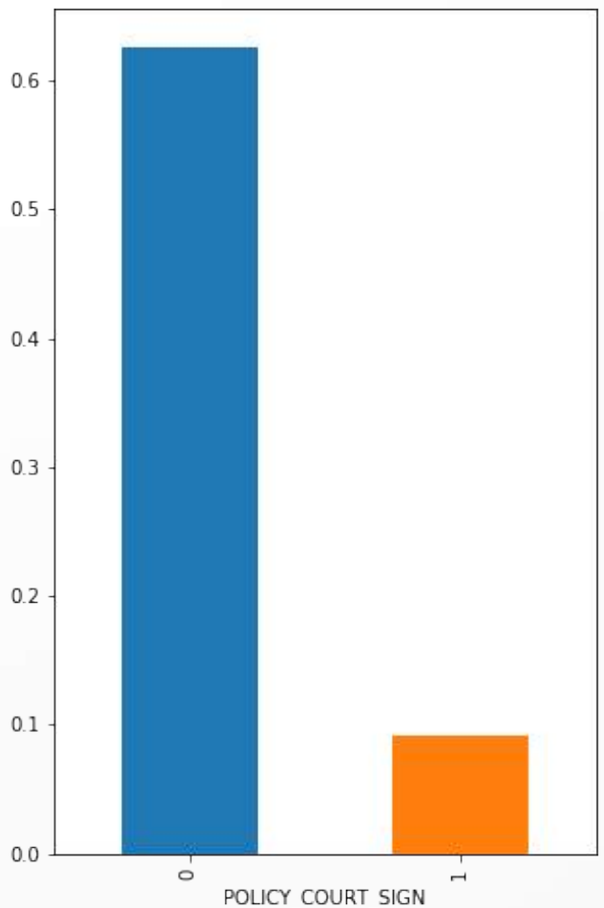
Вероятность продления полиса



Повышение цены полиса — отталкивающий фактор

Вероятность продления полиса





Жалобы и суды — отталкивающий фактор

Выводы

Из полученных данных можно сделать вывод, что вероятность пролонгации полиса выше у водителей старше 20 лет, имеющих стаж не менее 2 лет, уже продлевавших полис и не участвующих в ДТП. Отталкивающими факторами являются повышение цены полиса, жалобы и суды по полису.