Аналитическая записка.

Использованы различные модели для построения модели бинарной классификации.

Выбраны те модели, которые не склонны к переобучению из-за мультиколлинеарности, но в то же время интерпретируемы.

В исходном датасете разбалансированы классы.

Так, если случайно присваивать разбалансированные в соотношении 0,7:0,3 значения 0 и 1, то accuracy составит 0,5, а метрика Precision 0,3. В RandomForrest значение составило 0,73, что вкупе с Precision =0.37, что однозначно говорит о пользе модели.

В классификаторе AUC ROC = 0,75, и означает вероятность того, что случайно выбранный positive объект будет проранжирован классификатором выше (будет иметь более высокую вероятность быть positive), чем случайно выбранный negative объект.

CatBoost сработал лучше - Accuracy: 0.77, Precision: 0.42 и ROC AUC 0.81, что ещё выше случайных разбалансированных значений.

Однако, недостатком модели CatBoost являются более высокие требования к вычислительным мощностям.

Таким образом, если стоит задача интерпретируемости моделей, то есть смысл остановиться на CatBoost

В зависимости от того, какая стоит бизнес - задача, уменьшение ложноотрицательных или ложноположительных значений, есть смысл работать с показателем варьированием порога predict\_proba

Дальнейшая работа по улушению качества модели может двигаться исходя из создания новых призанков, так и из каких-либо дополнительных внутренних или внешних данных о клиентах (например, срок урегулирования страховых случаев, факт продления клиентом других страховок, продление других клиентов, относящихся к тому же кусту, скоринговой оценки БКИ или мобильных операторов, данных портала Госуслуг, ГИБДД, налоговой инспекции, данных приложений смартфона и проч).

В случае, если есть возможность отойти от требования интерпретации (например, если такое требование исходит не от регулятора, а от подразделения продаж или маркетинга), есть смысл попробовать нейросети, стекинг моделей, деформацию признаков. При этом стоит учесть, что неинтерпретируемые модели склонны к переобучению, нуждаются в регуляризации и их обучение стоит прекращать, когда на валидации качество перестаёт расти.

Примечание.

При построении модели автор исходил из предположения, что случаи, когда страховая сумма = 0, не являюстя ошибочными, а являются значением страховой суммы после её уменьшения в случае наступления страховых случаев (тотальные ДТП, угоны).