**Лабораторная работа №7**

**Классификация**

**(максимум – 11 баллов)**

По данной работе необходимо подготовить отчет в формате блокнота JupyterNotebook(.ipynb) либо в текстовом виде (.pdf). В отчете должны быть:

1) исходные коды;

2) результаты выполнения;

3) ответы на текстовые вопросы.

**Основнаячасть (6 баллов)**

В этом задании дан датасет с реальными данными одной строительной компании о бронированиях на покупку недвижимости.

Бронирование может привести к заключению договора на покупку, а может быть отменено. Для планирования своих финансовых показателей компания хочет прогнозировать, приведет ли бронирование к заключению договора или нет.

Таким образом, вам предстоит построить модель для бинарной классификации бронирований.

В этом задании будем работать только с жилой недвижимостью, поскольку именно по ней имеется больше всего данных.

1. Загрузите данные из файла "база.csv". Столбцы (признаки) имеют следующий смысл:
   1. УИД\_Брони – уникальный идентификатор брони (для классификации его использовать не надо – это просто id)
   2. ДатаБрони–дата установления брони
   3. ВремяБрони–время установления брони (GMT+4)
   4. ИсточникБрони–способ оформления брони: МП — мобильное приложение, ручная — офис продаж (бинарный признак)
   5. ВременнаяБронь–подтвержденная заявка на бронь(«Нет») или заявка на бронь не подтверждена отделом продаж («Да») (бинарный признак)
   6. **СледующийСтатус– если пусто или «В резерве» — статус еще не определен, «Продана» — договор оформлен, «Свободна» — бронь снята (отменена). Это и есть целевой признак, который нужно научиться предсказывать**
   7. Город–город проекта
   8. ВидПомещения–один из 4 видов: жилые помещения, нежилые помещения, кладовые, паркинг. Нас будут интересовать бронирования только со значением этого признака «жилые помещения».
   9. Тип–количество комнат для жилых помещений (для остальных видов неактуально)
   10. ПродаваемаяПлощадь–общая продаваемая площадь помещения
   11. Этаж–этаж расположения помещения
   12. СтоимостьНаДатуБрони–актуальная стоимость по прайсу на дату установления брони
   13. ТипСтоимости– «Стоимость при 100% оплате» или «Стоимость в рассрочку», остальные варианты можно считать отсутствием значения (бинарный признак)
   14. ВариантОплаты– «Единовременная оплата» или «Оплата в рассрочку» (бинарный признак)
   15. ВариантОплатыДоп–уточненный вариант оплаты: Ипотека/Вторичное жилье, если "пусто" — использовать ВариантОплаты
   16. СкидкаНаКвартиру–размер предоставленной скидки при её наличии, отрицательное значение — это наценка на помещение (как правило, при выводе из закрытого ассортимента)
   17. ФактическаяСтоимостьПомещения–стоимость заключения сделки (стоимость на дату брони за вычетом скидки)
   18. СделкаАН–участие в сделке агентства недвижимости (да/нет) (бинарный признак)
   19. ИнвестиционныйПродукт–да/нет — признак продажи помещения по инвестиционному договору (бинарный признак)
   20. Привилегия–да/нет — признак продажи помещения по инвестиционному договору типа Привилегия (бинарный признак)
   21. Статус лида (из CRM)–Статус клиента: S - успешный, P - в работе, F - забракованный
2. Предварительная фильтрация.
   1. Поскольку нас интересуют только сделки с жилой недвижимостью, отфильтруйте данные, оставив только те, для которых «ВидПомещения» = «жилые помещения». В дальнейшем этот столбец использоваться не будет, его можно удалить (или удалите его из датасета вообще, или просто нигде далее не рассматривайте).
   2. Также для нас бесполезны данные, по которым статус не определен. Отфильтруйте данные по признаку «СледующийСтатус». В оставшихся строчках замените значение «Продана» на 1, «Свободна» – на 0.
   3. Не забывайте, что столбец «УИД\_Брони» для нас также не представляет интереса – удалите его из датасета вообще, или просто нигде далее не рассматривайте.
3. Проверьте тип данныхи преобразуйте все данные к числовому типу.
   1. Для тех полей, которые по смыслу являются числовыми (например, «ПродаваемаяПлощадь») – просто проверьте правильность типа.
   2. Для бинарных признаков (например, «ИсточникБрони») выполните кодирование (один вариант закодируйте 0, другой 1).
   3. Для категориальных не бинарных признаков (например, «Город») выполните one-hotкодирование.
   4. Обратите внимание на поле «Тип». По смыслу оно числовое (количество комнат), но напрямую сконвертировать его в числовой тип мешает буковка «к» в конце. Напишите вручную преобразование, которое удаляет букву «к» в конце и конвертирует то, что осталось, в число. Если это невозможно (среди данных вам встретится еще вариант, когда в этом поле записано просто «с») – просто пока оставьте поле пустым (NaN).
4. Проверьте, есть ли по каким-либо признакам отсутствующие данные.
   1. Отсутствующие данные в поле «СкидкаНаКвартиру» замените на 0 (это значение по умолчанию – если поле не заполнено, то скидки, по всей видимости, нет).
   2. Отсутствующие данные в полях «Тип» и «ПродаваемаяПлощадь» замените на медианное значение, вычисленное по всему набору данных (*признаки кажутся достаточно важными, поэтому удалять эти столбцы не хочется; пустых значений довольно много, поэтому удалять строки тоже не очень хорошо; какого-то значения «по умолчанию» для этих полей нет; поэтому заменить эти значения на медиану представляется наилучшим решением*).
   3. Что делать с полем «ВариантОплатыДоп» решите самостоятельно (можно, как указано в описании, вместо пустых значений использовать значение из поля «ВариантОплаты», но в таком случае обратите внимание, что признак становится не бинарным; допустимо также совсем убрать этот столбец из рассмотрения).
   4. По всем остальным полям примите решение самостоятельно. Если отсутствующих данных не много, то удалите соответствующие строки.
5. Дополнение данных.
   1. Добавьте новый признак «Цена за квадратный метр». Он должен вычисляться на основе значений признаков «ФактическаяСтоимостьПомещения» и «ПродаваемаяПлощадь».
   2. Добавьте новый признак «Скидка в процентах», на основе значений «ФактическаяСтоимостьПомещения» и «СкидкаНаКвартиру».

*Комментарий. Многие алгоритмы классификации, по сути, учитывают только линейную зависимость. Поэтому, несмотря на то что добавляемые признаки полностью определяются значениями других признаков, их добавление имеет смысл – их зависимость от других признаков не линейная.*

1. Выполните нормализацию. Можете самостоятельно выбрать способ нормализации. «По умолчанию» предлагается выполнить минимаксную нормализацию и привести все значения к диапазону [0;1], кроме признака «СкидкаНаКвартиру» - его логичнее приводить к диапазону [-0,5; 0,5].
2. Проверьте датасет на сбалансированность (количество строк со значением целевого признака 0 и со значением 1). Является ли датасет сбалансированным?
3. Сформируйте список факторных признаков и целевой признак.
4. Выполните разбиение датасета на обучающую и тестовую выборки. При формировании обучающей и тестовой выборок строки из исходного датафрейма должны выбираться в случайном порядке.
5. Из библиотеки sklearn.neighbors возьмите алгоритм классификации KNN (KNeighborsClassifier). Постройте (обучите) модель. Для параметров используйте значения по умолчанию.
6. Из библиотеки sklearn.tree возьмите алгоритм классификации на основе деревьев решений (DecisionTreeClassifier). Постройте (обучите) модель. Для параметров используйте значения по умолчанию.
7. Получите векторы прогнозных значений целевой переменной на обучающей и на тестовой выборках для каждой из моделей.
8. Посчитайте показатели качества: «F-мера», точность (Precision) и полнота (Recall) на обучающей и на тестовой выборках для каждой из моделей.
9. Сделайте вывод о том, насколько хорошо удалось решить задачу прогнозирования. Какая модель оказалась лучше? Дайте интерпретацию полученных значений Precisionи Recall.

**Дополнительныезадания (5 баллов)**

1. (1 балл)Постройте boxplot («ящик с усами») для всех **числовых** признаков. Отфильтруйте исходные данные, удалив выбросы. Пересоздайте тестовую и обучающую выборки, переобучите модели.Посчитайте показатели качества. Как они изменились?
2. (2 балла) Выполните подбор параметров для алгоритмов KNN и деревьев решений. Для KNN попробуйте изменять параметр k – количество соседей, для деревьев решений – глубину дерева. Постройте графики зависимости показателей качества от значения параметра (от k в случае KNN и от глубины дерева в случае деревьев решений). Для параметра k рассматривайте диапазон от 1 до 40. Для глубины дерева – от 2 до 40. По графикам определите оптимальные значения параметров.
3. (1 балл) Из библиотеки sklearn.linear\_model возьмите алгоритм логистической регрессии (LogisticRegression). Постройте (обучите) модель. Посчитайте показатели качества. Сравните результат с другими моделями.
4. (1 балл) Из библиотеки sklearn.svm возьмите алгоритм SVM (машины опорных векторов) (LinearSVC). Постройте (обучите) модель. Посчитайте показатели качества. Сравните результат с другими моделями.