Автомобильная компания планирует выйти на новые рынки сбыта, на основе имеющейся базы клиентов в которой все клиенты разбиты на 4 группы требуется выполнить:

1. Оценить имеющуюся разбивку клиентов на группы
2. При необходимости скорректировать сегментацию рынка
3. Обучить модель для классификации новых клиентов

База состоит из 8000 клиентов классифицированных по 9-ти признакам. В исходных данных выбросы отсутствуют. 4 целевых категории разделены равномерно, что видно на диаграмме. Данные представляют собой в основной массе категориальные данные, такие как пол, наличие высшего образования или размер семьи. Данные содержат большое количество пропусков, от которых необходимо избавиться, что на данном этапе стало первоочередной задачей.

Стратегия обработки пропусков шла по 4-ем направлениям:

1. Заполнение пропусков с предварительной группировкой и удалением дубликатов
2. Уделение строк с пропусками
3. Удаление срок с пропусками и удаление дубликатов
4. Удаление срок с пропусками, удаление дубликатов и удаление противоречивых данных, это данные которые смеют одинаковые факторные признаки и разные результативные.

На диаграмме можно увидеть 4 результативных признака и изменение их числа в зависимости от метода обработки, по диаграмме можно сделать вывод, что объемы признаков остаются в любом случае равномерно распределены и их объем достаточен для обучения моделей.

Какие же данные лучше взять для работы мы определим путем обучения модели случайного леса. Эта модель была выбран на основе предварительного обучения различных моделей как хорошо себя зарекомендовавшая.

На диаграмме видна зависимость доли правильных ответов модели от способа обработки пропусков и по ней можно выделить, что метод 2 дает наиболее точные данные, а это метод заполнения пропусков, данные обработанные на основе этого метода далее и будем использовать.

Отбор признаков.

Данная задача решалась путем комбинирования модели случайного леса и тремя способами отбора признаков, это метод selectkbest, рекурсивное исключение и метод главных компонент. По диаграмме можно заметить, что качество модели начинает улучшаться начиная с 6 признаков и выше.

Важными признаками оказались: «Возраст», «Семейное\_положение», «Образование»

Несущественными: «Пол», затем «Опыт\_работы» и «Категория\_товара».

Обработку данных будем выполнять на всех 9-ти признаках, но будем иметь в виду, что и 6, и 7, и 8 тоже хорошо.

В заключении данного блока хотелось бы еще обратиться к корреляционной матрице из которой можно увидеть, что корреляции всех признаков с результативным очень низкие, не превышают 0.23, а из этого можно сделать предположение, что модели будут обучаться не очень охотно.

Кластерный анализ.

Целью было подтвердить, что на основе имеющихся признаков клиенты классифицированы верно, в случае удачи кластеры и результативные признаки должны частично или полностью совпадать, в обратном случае возникнут сомнения в рациональности исходной разбивки клиентов.

Кластеризация выполнялся на основе метода k средних путем разбивки на 4 группы на основе 9-ти признаков, результаты можно увидеть на круговых диаграммах.

Мы видим 4 результативных признака окрашенные в разные цвета, а это идеальное совпадение кластеров и признаков, т.е. каждому кластеру соответствует свой признак. Обращаясь к левой группе мы видим, что у нас получилось по факту, а именно, что в первый признак входит 4 кластера в разной пропорции, во второй тоже и т.д., т.е. можно сделать вывод , что данные подсказывают нам иной способ разбивки, отличный от разбивки, которая была выполнена изначально.

Было сделано предположение , что совпадение будет явным с уменьшением числа признаков, но результат не изменился.

Если на основе корреляционной матрицы можно увидеть слабые связи отдельных признаков с результативным, то здесь можно говорить о слабой связи всех признаков в совокупности с результативным, что может говорить о слабой предсказательности построенных моделей.

Дополнительно был выполнен иерархический кластерный анализ, с целью узнать правильно ли было выбрано количество классов исходной классификации. После построения дендрограммы выяснилось, что разбивка на 2, 3, 4 группы вполне уместна и наша исходная разбивка на 4 класса не противоречит полученным результатам.

Показала, что причисление клиента к той или иной группе в исходных данных не совсем рациональна и разбивка на 4 группы вполне допустима.

Переходим к обучению моделей

Были обучены ряд моделей: линейные, нелинейные и ансамбливые

Модели обучались для начала с параметрами заданными по умолчанию путем перебора самих моделей и перебора числа признаков методом selectKBest, результаты были представлены в графическом виде с помощью matplotlib. На диаграмме показаны зависимости качества моделей от числа признаков, т.е. по оси Х показано число признаков, а по Y среднее число правильных ответов. Можно заметить, что скачек качества моделей наблюдается при 6-ти признаках. Лучше всего себя проявила модель gradientboostingclassifier, которая достигает 53%-го числа правильных ответов, если подвести итог, то можно заключить, предсказательные возможности моделей остаются на низком уровне, поэтому было решено выполнить перебор параметров моделей с помощью метода gridseachCV.

В результате были получены данные, которые должны обеспечить наилучшее качество, результаты также были отображены в графическом фиде по аналогии с предыдущей диаграммой, т.е. показана зависимость качества моделей от числа признаков.

На диаграмме видно, что качество моделей подросло, некоторые подтянулись к лидеру на предыдущем графике gradientboostingclassifier, но лидер остался тот же, максимальный результат которого получилось добиться – это 55% правильных ответов, что также не может нас устраивать.

Классифицировать новых клиентов по такой неправдоподобной модели не очень хорошо, т.к. делать какието выводы не сумев правильно классифицировать клиентов = это неправильно.

Поэтому было решено обратиться к результатам кластеризации, выполнить анализ полученных результатов, обучить модель и дать рекомендации по дальней действиям.

Напомню, что иерархическая кластеризация подтвердила возможность разбивки на 4 признака, поэтому, если мы разобьем данные на 4 кластера по аналогии с исходными данными, это не будет противоречием.

По результатам кластеризации методом k-средних я вывел в таблицу для каждого кластера средние значения по признакам и вывел корреляционную матрицу. На основе корреляционной матрицы был выделен самый влиятельный признак – это возраст. В первой таблице для столбца – возраст была определена четкая разбивка по возрастам, на основе чего были даны названия кластерам:

Кластер 1 – Зрелый возраст (≈ 39 лет)

Кластер 2 – Молодежь (≈ 26 лет)

Кластер 3 – Пожилые люди (≈ 75 лет)

Кластер 4 – Средний возраст (≈ 52 года)

Таким образом мы получили более рациональную разбивку данных, лучшую корреляцию факторных признаков и результативных при прежней слабой корреляции факторных признаков между собой, что предположительно должно позволить обучить более качественную модель.

Для наилучшего результата обучение производилось с помощью зарекомендовавшей себя gradientboostingclassifier. Результаты были оценены с помощью метода classification\_report, confusion\_matrix и путем вывода доли правильных ответов тестовой выборки.

В итоге был достигнут практически 100% результат. Из первого метода можно увидеть, что только первый признак дал 99%, остальные 100%. В confusion\_matrix показал почти 100%-е совпадение и средняя доля правильных ответов тоже выдала 100% результат.

Вывод:

1. Классификация исходных данных некорретна
2. Разбивка на 4 класса возможно
3. Главное возраст(как в старой так и в новой классификации)
4. Модели на исходных данных дают низкую точность – 55%
5. Модели на кластеризированных данных дают высокую точность 99%

В заключении можно предположить, что возможно разбивка была выполнена по другим признаках, либо главный признак классификации не вошел в датасет.

Поэтому хотелось бы рекомендовать в качестве классификации новых клиентов использовать результаты кластерного анализа, т.к. это приводит к тому, что клиент будет со 100%-ой вероятностью правильно причислены к той или иной группе, на основе безошибочной классификации можно будет уверенно проводить анализ можно, делать прогнозы и рекомендовать тот или иной товар, не опасаясь, что новый клиент попал не в ту группу.