Московский Государственный Университет им. М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра исследования операций

Приходько М. Н.

Исследование рынка ОСАГО РФ за 2019 г. с применением алгоритма линейной регрессии.

Курсовая работа

Научный руководитель: к. ф.-м. н. Белянкин Г.А.
Со-руководитель: к. ф.-м. н. Чулков С.П.

Москва, 2021

Содержание

Вступление

В данной работе проводится исследование рынка ОСАГО[1] в России за 2019 год с применением статистического анализа данных. Проводится исследование с использованием статистических данных из различных источников и на их основании строится модель. В ходе работы с данными будет изучена модель линейной регрессии. Основной задачей при анализе данных и построении модели будет выявление закономерностей между показателями работы субъектов страхового дела (далее: ССД[2]). Исследование проводилось в Jupyter Notebook — основная программа на Python (см. Приложение №4).

1 ОСАГО - обязательное страхование гражданской ответственности владельцев транспортных средств

2 ССД — субъект страхового дела. В рассматриваемом исследовании — Страховая по ОСАГО.

Постановка задачи

Ситуация и гипотеза.

Существует ситуация, когда компании-страховой необходимо узнать от каких параметров её деятельности зависит частота обращения автовладельцев в суд. Это может быть полезно для данной страховой для оптимизации своей работы. Для этого формируется **гипотеза**, что оценка частоты :

$$frequency = (\frac{\text{Количество судебных дел}}{\text{Количество полисов ОСАГО}})$$

в разрезе Страховых компаний и Регионов может объяснятся следующими показателями этого ССД:

1) Размер средней выплаты
2) Размер средней страховой премии
3) Региона действия

Задачи исследования

- Сбор статистической информации за 2019 год по ССД, содержащей описанные выше показатели. Объединение указанных данных.
- Проверка **гипотезы** путём построение модели линейной регрессии на параметрах 1) — 3) для предсказания целевого значения frequency.
- По оценки модели на тестовой выборке подтвердить или опровергнуть.

Линейная регрессия (Linear regression)

Определение

- **Регрессия (Regressions)** предсказывает значения выходных переменных, основываясь на численных значениях входных переменных.
- **Линейная регрессия** - один из наиболее хорошо изученных методов машинного обучения, позволяющий прогнозировать значения количественного признака в виде линейной комбинации прочих признаков с параметрами - *весами* модели. Оптимальные (в смысле минимальности некоторого *функционала ошибки*) параметры линейной регрессии можно найти аналитически с помощью нормального уравнения или численно с помощью методов оптимизации.

Исторический обзор

Модель линейной регрессии предполагает, что функция регрессии $E(Y \mid X)$ линейна на входах $X_1,..., X_d$. Линейные модели, как целый класс, были разработаны в основном в докомпьютерную эпоху статистики, но даже сегодня, в цифровую эпоху по-прежнему остаются веские причины для их изучения и использования. Они просты и часто дают адекватное и поддающееся интерпретации описание того, как входные данные влияют на выходные данные. Для целей прогнозирования они иногда могут превосходить более сложные нелинейные модели.

Описание линейной модели

В обычной ситуации мы имеем набор данных для обучения модели $(x_1 , y_1)…(x_n , y_n)$, который необходим для поиска и оценки *весов модели* w. Каждый $x_i = (x_{i1}, x_{i2}, …, x_{in})^T$ - это вектор измеримых признаков на i-ом наблюдении. Признак - это число, характеризующее объект.

Линейный алгоритм в задачах регрессии выглядит следующим образом:

$$a(x) = w_0 + \sum_{j=1}^{d} w_j x^j$$

где a(x) — *выбранный алгоритм*, w_0 — *свободный коэффициент*,

x_j - *признаки*, а w_j — *их веса*.

Если добавить (d + 1)-й признак, который будет на каждом объекте принимает значение 1, *линейный алгоритм* можно будет записать в более компактной форме:

$$a(x) = \sum_{j=1}^{d+1} w_j x^j = \langle w, x \rangle$$

где для скалярного произведения двух векторов используется обозначение (w, x).

Оценка модели. Мера ошибки

Для оценки точности предсказаний модели обычно используют различные меры ошибок.
Простое отклонение от прогноза:
Q(a, y) = a(x) − y нельзя использовать, поскольку минимум функционала не будет достигаться при правильном ответе a(x) = y.

Одним из вариантов меры ошибки можно считать:
Q(a, y) = |a(x) − y| .
Но этот функционал не обладает нужными для оптимизации свойствами гладкости.

Самым применимым считается квадрат отклонения:
Q(a, y) = (a(x) − y)².

На его основе вычисляется функционал ошибки, называемый среднеквадратической ошибкой алгоритма (MSE3).

2)

$$Q(a, x) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 \qquad Q(w, x) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2$$

3 MSE (англ.) - Mean Squared Error

1) Среднеквадратическая ошибка алгоритма (общий случай).

2) Среднеквадратическая ошибка алгоритма (для линейной регрессии).

Среднеквадратическая ошибка для линейной регрессии, как функционал, обладает необходимым свойством гладкости, которое позволяет применить к нему известные методы оптимизации.

Задача линейной регрессии

Задача обучения алгоритма линейной регрессии — задача минимизации функционала ошибок, как правило MSE. Решая такую оптимизационную задачу, мы приближаем предсказания линейной модели к самим наблюдениям целевого признака. В общем виде задача выглядит следующим образом:

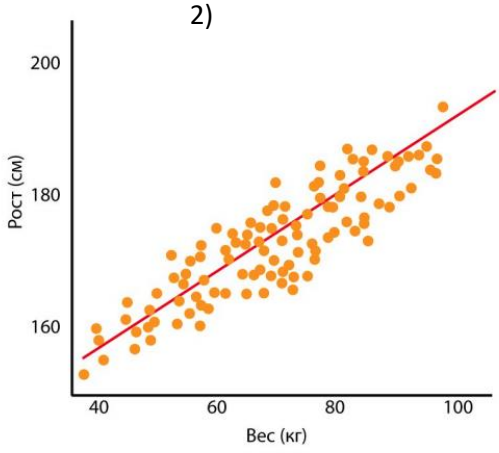
$$Q(w, x) = \frac{1}{\ell}\sum_{i=1}^{\ell}(\langle w, x_i\rangle - y_i)^2 \to \min_w .$$

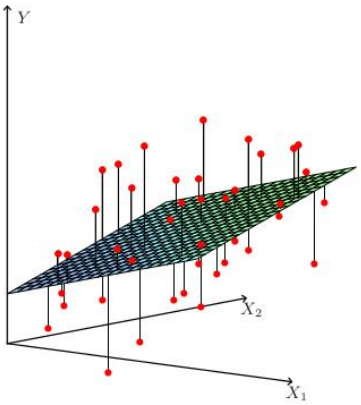
Возможно переписать с помощью матричных обозначений, если рассматривать матрицу признаков *X* и вектор ответов *y :*

$$Q(w, X) = \frac{1}{\ell}\|Xw - y\|^2 \to \min_w$$

Геометрическая интерпретация

2)





Примеры геометрической интерпретации задач линейной регрессии. Гиперплоскость, которая получается в результате работы метода (прямая для 1-мерного случая, плоскость для 2-мерного и т.д.), строится таким образом, чтобы минимизировать установленный для данной задачи функционал ошибки (как правило MSE). Минимизируются расстояния (метрические проекции см. Приложение №1) от целевых объектов до построенной гиперплоскости и усредняется по всем объектам.

Оценка результатов работы модели линейной регрессии

1. MSE - **среднеквадратичная ошибка**.

Напомним, что для линейной регрессии формула имеет вид:

$$Q(w,x) = \frac{1}{\ell}\sum_{i=1}^{\ell}(\langle w, x_i\rangle - y_i)^2$$

где *l* — число наблюдений, (w, x_i) — предсказанное моделью значение для y_i , y_i – истинное целевое значение. Чем выше значение MSE, тем больше общая квадратичная оценка, и тем хуже модель. Помимо отмеченного выше преимущества в виде гладкости, полезной для оптимизации, возведение ошибки в квадрат несколько крупных ошибок больше, чем множество мелких.

2. R^2 - **коэффициент детерминации**.

R^2 - измеряет величину дисперсии в векторе целей, которая объяснятся моделью:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Немного раскроем смысл использование такой оценки качества обучения алгоритма линейной регрессии.

В ходе оценки регрессионных моделей R^2 интерпретируется как соответствие модели данным.

- Для приемлемых моделей предполагается, что коэффициент детерминации должен быть $\geq$ 50% (в этом случае коэффициент множественной корреляции превышает по модулю 70%).
- Модели с коэффициентом детерминации выше 80% можно признать хорошими (коэффициент корреляции превышает 90%).
- Значение коэффициента детерминации 1 говорит о функциональной зависимости между переменными.

Используемые статистические данные

Для проводимого исследования было задействовано несколько источников статистических данных.

Первым источником статистики послужила информация о судебных делах, связанных с исками по ОСАГО. Для необходимого анализа с помощью написанной программы на Python (листинг программы — см. Приложение №2), была собрана информация о судебных делах, удовлетворяющая следующим критериям:

- Ответчик по судебному делу — компания страховая
- Рассматриваемый период судебного дела — 2019 г.
- Территория суда, где рассматривалось дело, удовлетворяет одному из 8 регионов РФ, список приведен в программе на Python (листинг программы — см. Приложение №4)
- Страховая — ответчик является ССД.

В ходе исследования информация по делам была сгруппирована по страховым и по регионам. Таким образом была получена информация о количестве судебных дел в разрезе компаний - страховщиков по регионам РФ.

Вторым, важным источником статистических данных послужила официальная отчетность по ключевым показателям ЦБ РФ по ССД за 2019 год.

Из этих статистических данных

- Сведения о суммарные выплаты по ССД (1)
- Сведения о количестве урегулированных дел по ССД (2)
- Сведения о количестве договоров ОСАГО по ССД (3)
- Сведения о суммарных страховых премиях по ССД (4)

С помощью программы-скрипта на Python (листинг программы — см. Приложение №3) были получены данные о средних премиях и средних выплатах по ССД, которые далее использовались как параметры регрессионной модели.

Ход исследования

Получение данных

О данном этапе исследования достаточно подробно рассказано в предыдущем разделе.

Поэтому остановимся немного подробнее на технических деталях получения данных. Для получения и сбора данных по судебным делам в программе (листинг программы — см. Приложение №2) применялся web-parsing.

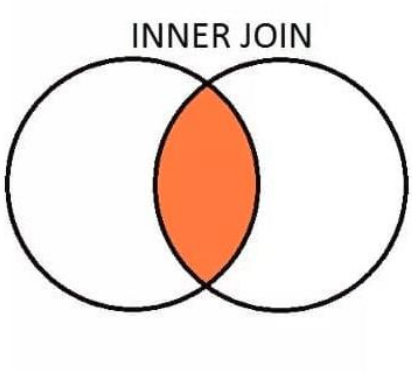
Web-parsing — анализ разметки HTML и связанный с этим процесс извлечения данных, которые нам необходимы.

Web-parsing был реализован использованием библиотеки Python BeautifulSoup4. Интересным моментом текущего этапа стало время работы программы, затрачиваемое на сбор данных по одному региону за 2019 календарный год. Так например для Ростовской области для сбора данных с порядка 365 * 50 = 18250 интернет страниц было затрачено порядка 19000 секунд работы алгоритма, что составляет порядка 5 часов непрерывной работы программы (365 — дней в году, 50 — примерное число районных судов в Ростовской области). Для хранения промежуточных данных по регионам были использованы таблицы Excel.

Статистические, отчетные данные ЦБ РФ были скачаны с официального сайта ЦБ РФ (см. Приложение №5).

Предобработка данных

Для необходимого анализа данных были объединены таблицы со статистическими данными от ЦБ РФ (см. Приложение №3) с использованием метода merge из библиотеки Pandas с помощью способа Inner Join, что означает пересечение данных по равенству ключевых признаков, схематическое изображение модели объединения:



Были унифицированы названия для страховых из собранной информации по судебным делам.

Далее были объединены данные по судебным делам и ключевые показатели ССД. В результате чего была получена сводная таблица.

Анализ данных

После создания сводной таблицы и применении метода value_counts() из Pandas Python были получены интересные для анализа данные о количестве судебных дел для каждой из страховой в рассматриваемых 8 регионах за 2019 год:

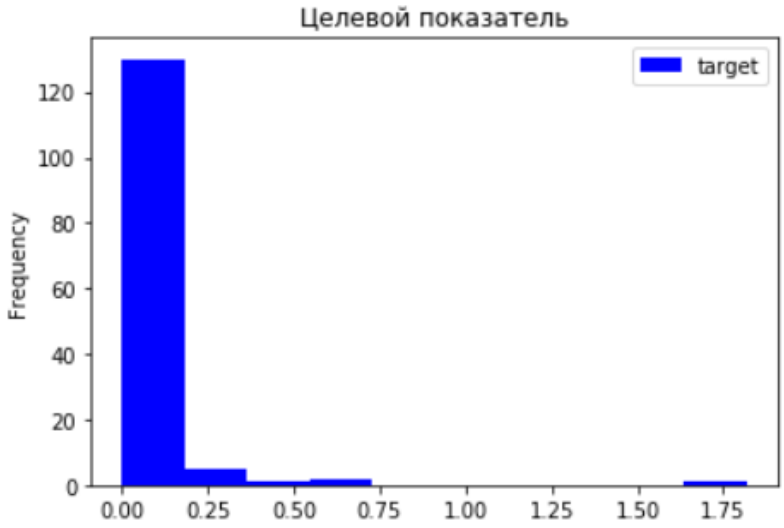
1. ПАО СК "Росгосстрах" 9772
2. АО "АльфаСтрахование" 3916
3. СПАО "РЕСО-Гарантия" 3794
4. СПАО "Ингосстрах" 3613
5. АО "СОГАЗ" 2507
6. САО "ВСК" 2363
7. АО "Группа Ренессанс Страхование" 1758
8. АО "МАКС" 1196
9. Либерти Страхование (АО) 346
10. САО ЭРГО 341
11. ПАО "АСКО-СТРАХОВАНИЕ" 338
12. АО "Тинькофф Страхование" 277
13. ООО "Зетта Страхование" 125
14. ООО "СК "Согласие" 111
15. ООО "СФ "Адонис" 99
16. ООО СК "Паритет-СК" 67
17. ООО "Абсолют Страхование" 40
18. ПАО "САК "ЭНЕРГОГАРАНТ" 36
19. ООО "СК "Мегарусс-Д" 23
20. АО "Страховая бизнес группа" 17
21. САО "Медэкспресс" 10
22. ООО СО "Геополис" 8
23. АО "Страховая Компания "ПОЛИС-ГАРАНТ" 6

Таким образом видно, что весомую долю судебных дел имеет ПАО СК "Росгосстрах" (9772). Объединенная сводная таблица с ключевыми показателями уже содержит все нужные параметры регрессии и значения целевого признака. В результате получилась выборка из 139 наблюдений, что, конечно, мало для построения регрессионной модели.

Обучение регрессионной модели. Анализ взаимосвязи ключевых параметров ССД.

О работе алгоритма линейной регрессии подробно рассказывается в одной из написанных раннее глав.

Перед построением непосредственно модели была подготовлена визуализация распределения целевой frequency:



Таким образом видно, что в большинстве случаев целевая частота имеет значение $0 \leq$ frequency ≤ 0.25.

Были изучены зависимость параметров *целевой частоты* от *средней выплаты за 2019 г. по ОСАГО:*

Также получена и визуализирована зависимость *целевой частоты* от *средней премии за 2019 г. по ОСАГО:*

Для модели, относительно начального рассмотрения, были добавлены 3 искусственных параметра:

1. *(Средняя выплата за 2019 г.)* ^{2}
2. *(Средняя премия за 2019 г.)* ^{2}
3. *(Средняя выплата за 2019 г.) * (Средняя премия за 2019 г.)*

Такой шаг связан с возможной нелинейной зависимостью целевого значения от входных параметров. Если между параметрами взаимосвязи нет, то получаемые веса у этих признаков будут почти 0.

Категориальный признак региона из исходных данных (один из 8 регионов) метода OneHotEncoder из sklearn.preprocessing был закодирован 8 бинарными признаками, которые равны 1, если текущая страховая в рассматриваемом регионе и 0 иначе, для остальных регионов.

Далее исходные данные при помощи train_test_split из sklearn.model_selection были поделены на обучение и тест для модели в отношении 9/1.

Далее были обучены регрессионные алгоритмы из библиотек statsmodels и Scikit Learn. И получены предсказания на тестовой выборке. Далее перейдем к оценкам получившихся моделей и итогам всего исследования.

Итог исследования. Выводы

Оценка на тесте

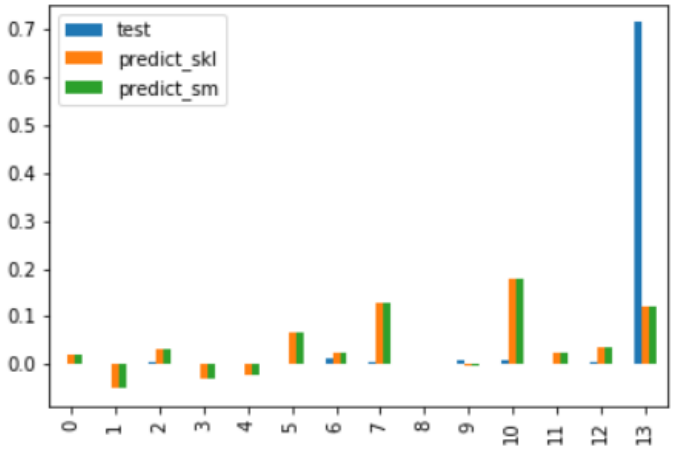
Модели регрессии из Statsmodels и Scikit Learn показали одинаковые предсказания и соответственно результаты MSE (средняя квадратичная ошибка) и MAE (средняя абсолютная ошибка).

MSE = 0.029191819643873167
MAE = 0.08471717662963665

В информации модели линейной регрессии Statsmodels доступна более подробна информация. Внимания заслуживает рассмотренная выше оценка качества:

R^2 =0.202

Построена сравнительная диаграмма результатов, предсказанных моделями значениями на тестовой выборке и значения целевого параметра частоты из теста:



Поскольку имеется такая интерпретация оценки «Для приемлемых моделей предполагается, что коэффициент детерминации должен быть $\geq$ 50%» и на основании приведенной выше столбчатой диаграммы можно заключить что модель плохо обучилась.

Вывод

Низкие результаты R^2 показывают, что взаимосвязь между рассматриваемыми параметрами:

1) Размер средней выплаты

2) Размер средней страховой премии

3) Региона действия

и

$$frequency = (\frac{\text{Количество судебных дел}}{\text{Количество полисов ОСАГО}})$$

частотой

не такая существенная. Поэтому гипотеза исследования не подтверждается.

Причины:

Не подтверждение гипотезы может быть связано с причинами:

- Маленькие данные для обучения (139 наблюдений).
- Использование Средней выплаты для каждой страховой без учета региона. Связано с отсутствием нужной статистике в аналитическом отчете ЦБ РФ.
- Возможно исследуемая взаимосвязь в действительности отсутствует.

Используемая литература

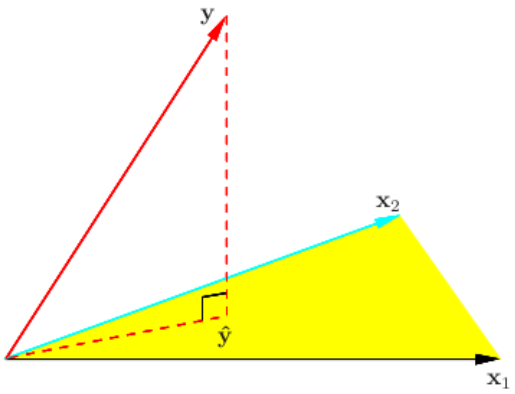
1) «The Elements of Statistical Learning» - Trevor Hastie, Robert Tibshirani, Jerome Friedman

2) «Big Data and AI Strategies Machine Learning and Alternative Data Approach to Investing» - Marko Kolanovic

3) «Машинное обучение с использованием Python. Сборник рецептов» - Крис Элбон

Приложение

1) Приложение №1: частный случай метрической проекции на множество — ортогональная проекция.



2) Osago_parser.ipynb

3)Analyse2019.ipynb

4)OSAGO_coursework.ipynb

5)http://cbr.ru/insurance/reporting_stat/

6) Данные по судам — all_data.xlsx

7) Статистические данные ЦБ РФ — CB_OSAGO_2019.xlsx