МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшегообразования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра дифференциальных уравнений и математической экономикимеханико-математического факультета

Решение задач классификации в пакете H2O
АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студента 247 группы

Направление – **09.04.03 Прикладная информатика**

механико-математического

факультета

Похазникова Ильи Алексеевича

Научный руководитель
профессор,
д.э.н., профессор _____ _____ В.А. Балаш

Заведующий кафедрой
зав. кафедрой, д.ф.-м.н.,
профессор _____ _____ С.И. Дудов

Саратов 2022

Введение. Благодаря машинному обучению программист не обязан писать инструкции, учитывающие все возможные проблемы и содержащие решения. При этом компьютер (или отдельное программное обеспечение) закладывает алгоритм самостоятельного поиска решений путем комплексного использования статистических данных, которые позволяют вывести закономерности и сделать прогнозы.

Машина машинного обучения на основе анализа данных начала свою историю в 1950 году, когда начали разрабатывать первые программы для игры шашками. За прошедшие десятилетия общий принцип не изменился. А потому из-за взрывного роста вычислительных мощностей компьютеров многократно усложнились закономерности и прогнозы, которые создаются ими в процессе машинного обучения.

Чтобы использовать машинного обучения, сначала вам нужно подгрузить нужный набор данных (определенный объем исходных данных) на компьютер, на котором алгоритм будет научится обрабатывать информацию. Например, разные фотографии, на которых есть метки, указывающие, кому они принадлежат. После процесса обучения программа сама без всякой помощи сможет распознавать метки на новом изображениях без содержания тегов. Обучение будет продолжаться даже после выдачи нужной информации, чем больше мы будет погравировать и анализировать с помощью программы, тем точнее она распознает нужные изображения.

Благодаря машинному обучению компьютеры учатся распознавать не только лица на фотографиях и рисунках, но и пейзажи, объекты, текст и цифры. Что касается текста, то здесь никак не обойтись без машинного обучения: функция проверки грамматики теперь присутствует в любом текстовом редакторе и даже в телефонах. Более того, учитывается не только написание слов, но и контекст, оттенки смысла и другие тонкие лингвистические аспекты. Более того, уже существует программное обеспечение, способное писать новостные статьи без вмешательства человека (на тему экономики и, например, спорта).

Цель и задачи практики

Данные относятся к случаю прямого маркетинга из страхового сектора, который пытался предсказать ответственность за политику. Речь идет о том, чтобы предсказать, кто будет заинтересован в покупке страхового полиса каравана.

Этот набор данных был использован во втором издании журнала Computational Intelligence and Learning соревновательный вызов, организованный CoIL cluster, который представляет собой сотрудничество между четырьмя финансируемыми ЕС. Сети передового опыта, которые представляют области нейронных сетей (NeuroNet), нечетких систем (ERUDITE), эволюционные вычисления (EvoNet) и машинное обучение, и он принадлежит и пожертвован Питером ван дер Путтеном.

Эти данные относятся к делу о прямом маркетинге из страхового сектора, который пытался предсказать ответственность за политику. Речь идет о том, чтобы предсказать, кто будет заинтересован в покупке страхового полиса каравана.

Основное содержание данной работы состоит из 4-х разделов, а именно:

1. Логистическая регрессия
2. Модели линейного дискриминантного анализа
3. Случайный лес
4. Бусстинг

Первый раздел рассмотрение логистическая регрессия, а именно:

Поскольку наша задача - предсказать, купит ли конкретный клиент полисы caravan, мы создаем матрицу путаницы.

```
probs <- predict(l.glm, type = "response")
pred.glm <- rep(0, length(probs))
pred.glm[probs > 0.5] <- 1
mobilehome = caravan$CARAVAN
confusionMatrix(table(pred.glm, mobilehome), positive='0')
```

```
Accuracy : 0.9401
Kappa : 0.0338
Sensitivity : 0.96854
```

```
            Specificity : 0.02011
             Prevalence : 0.94023
pred.glm.ROC <- roc(predictor=pred.glm, response=caravan$CARAVAN)
pred.glm.ROC$auc
plot(pred.glm.ROC, main="ROC")
```

Эта модель предсказала 15 клиентов, которые будут покупать наш полис, а 5807 не буду покупать полис. Из этих наблюдений следует, что чувствительность нашей модели составляет 96 %, то есть было правильно предсказано, что эти клиенты не будут покупать полис. Специфичность составляет 2 %. Это те клиенты, которые могут покупать наш полис и они естественно его купят.

Следовательно, эта модель обладает высокой чувствительностью, но очень низкой специфичностью.

Точность модели высока на уровне 94 %, ожидается, что набор данных сильно несбалансирован, только 341 (6 %) клиентов купили полисы мобильного дома. Если бы мы предсказали, что все клиенты не будут покупать полисы мобильного дома, мы все равно получим общий процент 5466/5822 правильных прогнозов, который составляет примерно 94 % (рисунок 16). В результате то, к чему мы стремимся с точки зрения ошибки обучения, - это модель с высокой специфичностью, поскольку мы пытаемся определить клиентов, которые будут покупать полисы мобильного дома.

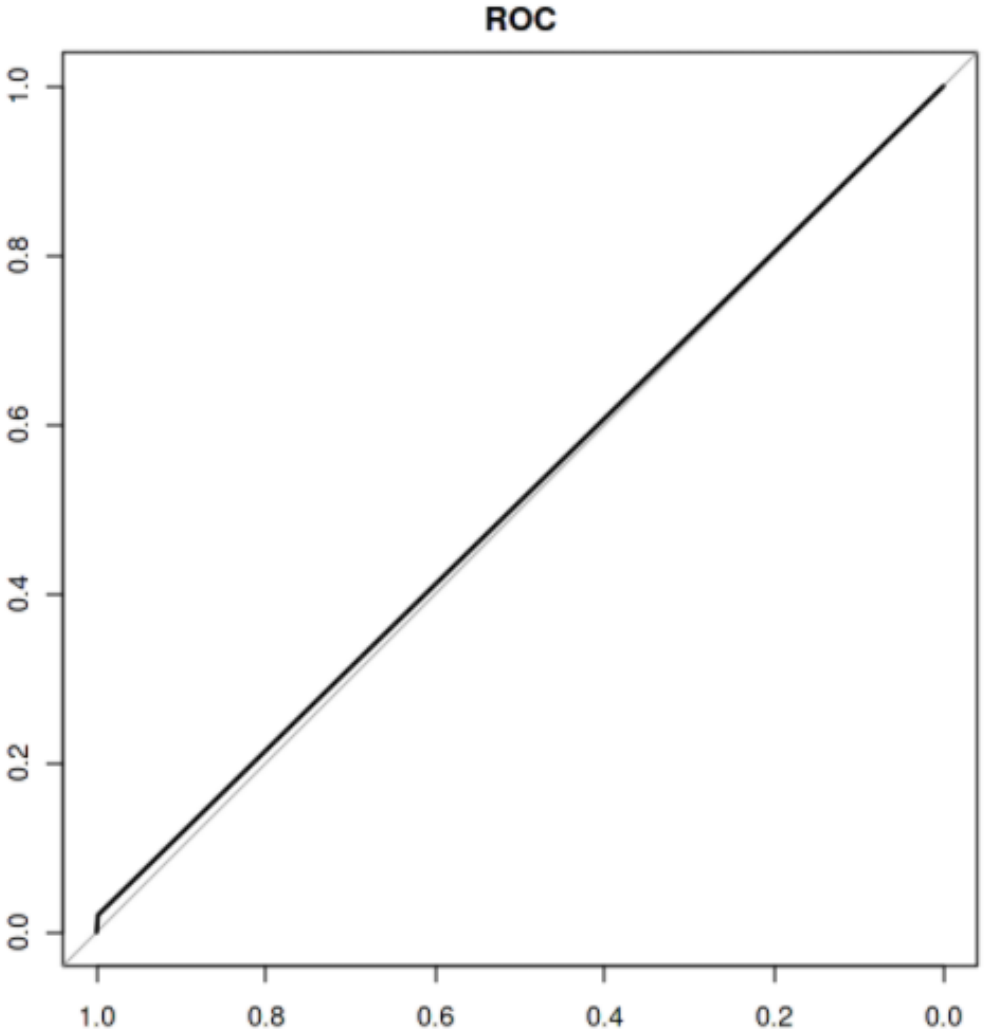

Рисунок 1 – ROC кривая метода логистическая регрессия

Для идеальной ROC кривой у хорошей модели, так это то, что чувствительность и специфичность должны быть близки к левому верхнему краю кривой, что означает, что площадь под кривой равна 1. Однако, как видно на этом графике, линия находится далеко от верхнего левого края, а площадь под кривой составляет всего 0,509, что также можно получить как сбалансированную точность в предыдущем разделе Матрицы путаницы и статистики (рисунок 1).

Второй раздел модель линейного дискриминанта, а именно:

Второй тип модели - это линейный дискриминантный анализ (LDA), который тесно связан с логистической регрессией, в которой оба производят линейные границы решения, которые отделяют класс от другого (рисунок 17). Единственное отличие заключается в том, что LDA будет предполагать, что наблюдения взяты из гауссовского распределения с общей ковариационной матрицей в каждом классе, и

если это предположение верно, оно будет работать лучше, чем регрессия Logisitc. [18]

```
fit.lda = lda(data=caravan, CARAVAN~.)
pred.lda = predict(fit.lda, type = "response")
confusionMatrix(table(pred.lda$class, caravan$CARAVAN), positive='0')
```

Accuracy : 0.9373

Kappa : 0.0914

Sensitivity : 0.95288

Specificity : 0.07022

Prevalence : 0.94523

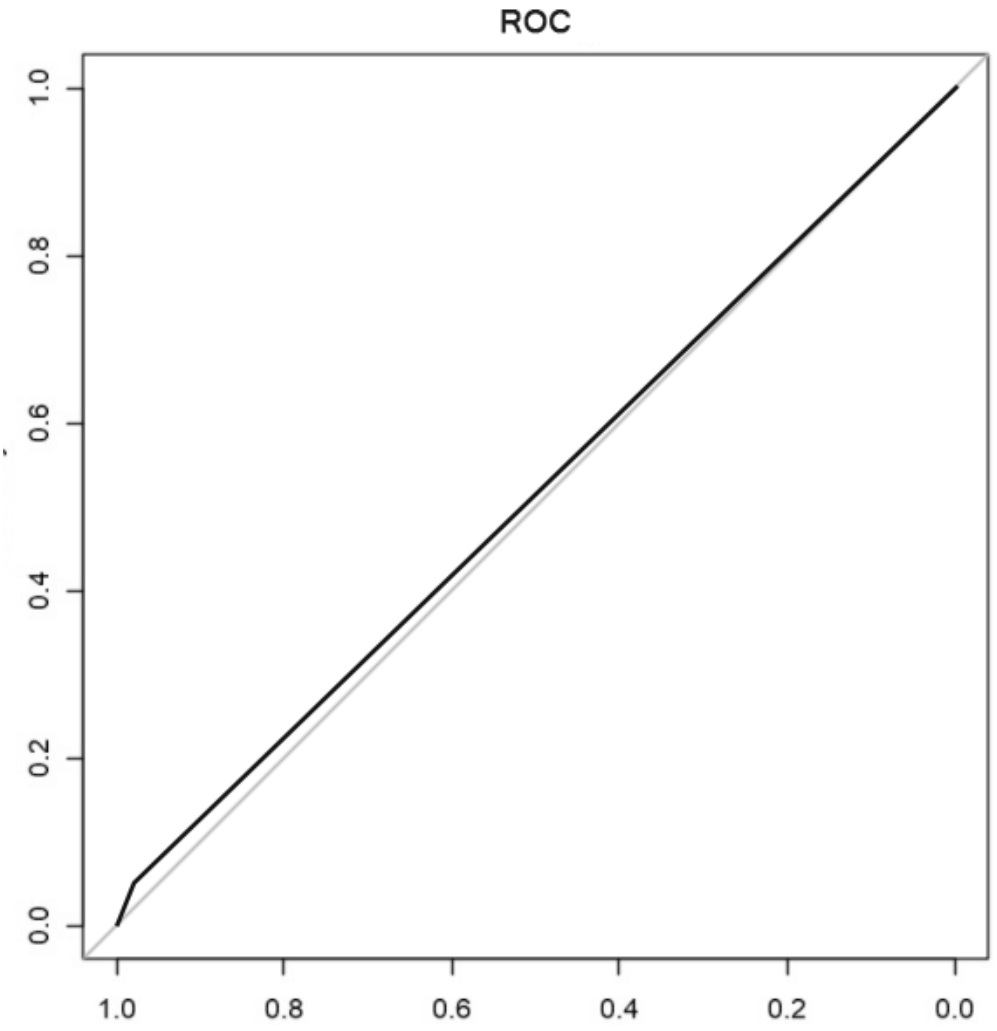

Рисунок 2 - ROC кривая метода линейный дискриминант

Чувствительность составляет 95 %, а специфичность – 7 %. Сравнивая модели линейного дискриминантного анализа с моделью логистической регрессии, чувствительность модели линейного дискриминантного анализа ниже, ее специфичность выше, чем 2 % в модели логистической регрессии. Если две или более переменных представляют собой почти линейную комбинацию друг с другом, их оценочные коэффициенты будут близки к 0, что затрудняет полную интерпретацию их влияния на целевую переменную. Стоит отметить, что мы должны избегать переменных, которые сильно коррелируют друг с другом (рисунок 2).

Третий раздел случайный лес, а именно:

Случайный лес H2O (RF) реализует распределенную версию стандарта алгоритм случайного леса и меры переменной важности. Сначала мы обучим базовую модель случайного леса с параметрами по умолчанию. Модель случайного леса выведет распределение ответов из кодировки ответа.

```
rf_fit1 <- h2o.randomForest(x = x,
                            y = y,
                            training_frame = train,
                            model_id = "rf_fit1",
                            seed = 1)
```

H2ORegressionMetrics: drf

MSE: 0.01183309

RMSE: 0.10878

MAE: 0.04327889

RMSLE: 0.07275499

Mean Residual Deviance : 0.01183309

AUC 0.631201482614

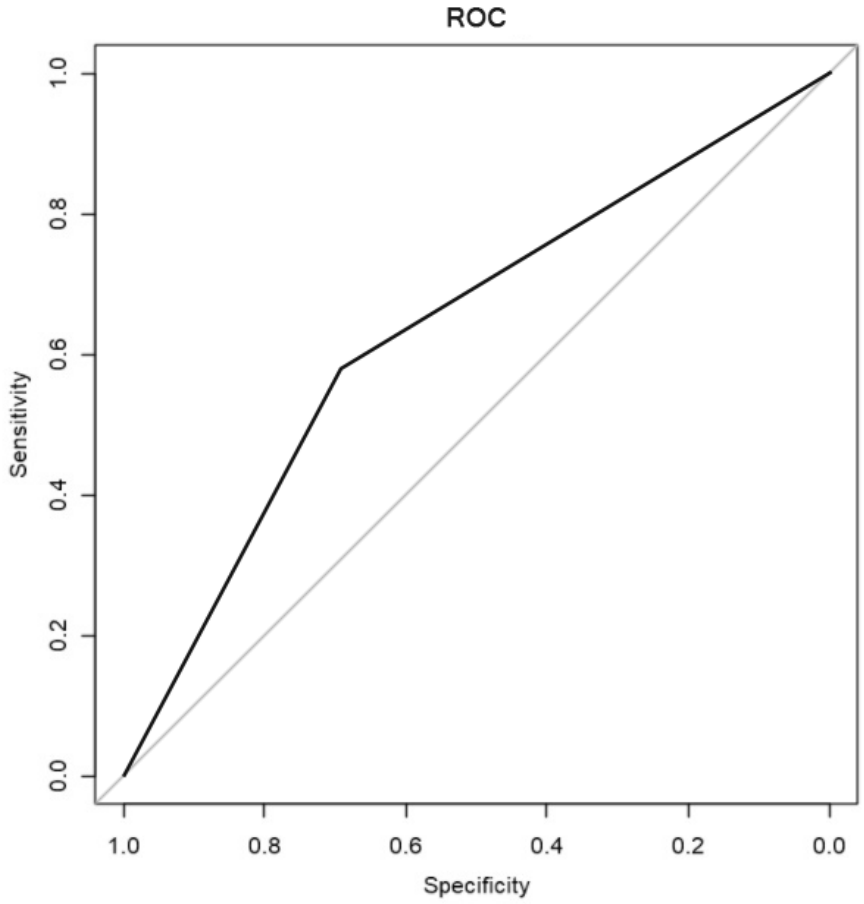

Рисунок 3 – случайный лес

Чувствительность составляет 96,5 %, а специфичность – 7,5 %. Сравнивая модели линейного дискриминантного анализа и моделью логистической регрессии, чувствительность модели линейного дискриминантного анализа ниже, специфичность выше, чем 6 % в обеих моделях.

Четвертый раздел метод бустинг, а именно:

Машина для повышения градиента H2O (GBM) предлагает стохастический GBM, который может значительно увеличьте производительность по сравнению с оригинальной реализацией GBM.

Теперь мы будем обучать базовую модель GBM

Модель GBM выведет распределение ответов из кодировки ответа, если она не указана явно через аргумент `distribution`. Для воспроизводимости требуется затравка (рисунок 3).

```
gbm_fit1 <- h2o.gbm(x = x,
                    y = y,
```

training_frame = train,

model_id = "gbm_fit1",

seed = 1)

MSE: 0.01283309

RMSE: 0.11178

MAE: 0.04727889

RMSLE: 0.08275499

Mean Residual Deviance : 0.01383309

AUC 0.661201482614

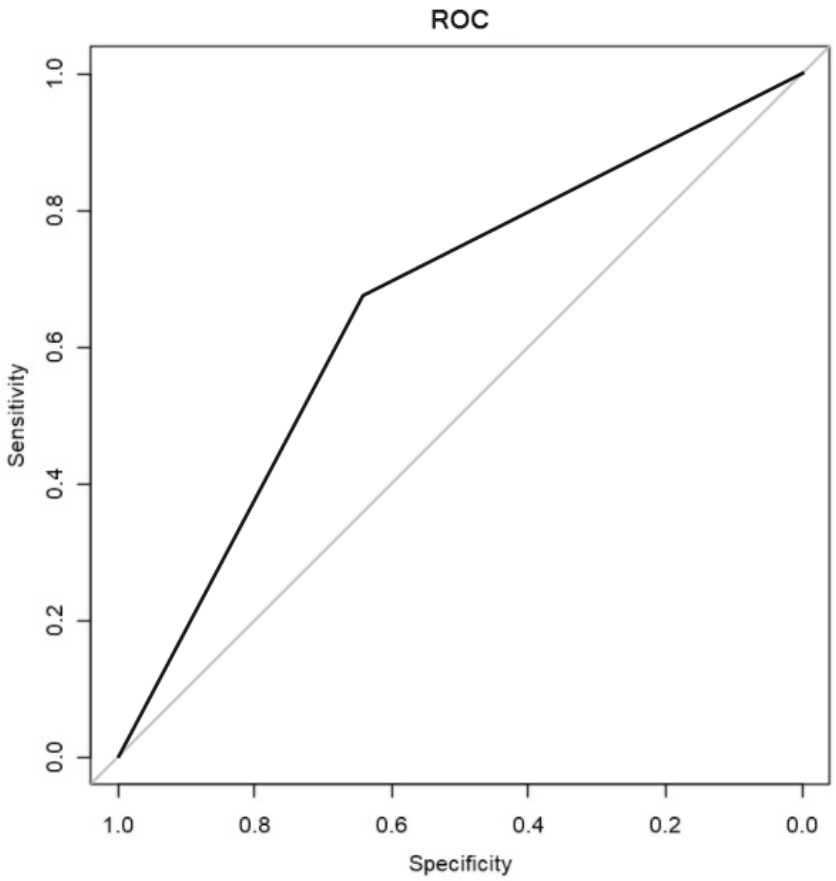

Рисунок 4– Roc кривая метода бустинг

Чувствительность составляет 97%, а специфичность – 7,7 %. Сравнивая модели линейного дискриминантного анализа и моделью логистической регрессии и случайного, чувствительность модели линейного дискриминантного анализа ниже, специфичность выше, чем 6 % в обеих моделях. Это лучшая из моделей машинного

обучения, он показывает лучшие результаты, это наталкивает на мысль, что её надо совершенствовать дальше и обучать (рисунок 4).

Заключение. Основные результаты, полученные в ходе магистерской работы:

Машинное обучение и нейросети всё больше проникают в отрасль науки, технологий и бизнеса. Они могут решать самые сложные задачи, распознавание лиц, поиск людей, анализы медицинского характера, биджевых рынков, подбор музыки, которая вам может понравиться. Машинное обучение — это не только обучение в общем плане. Даже глубокое обучение не может позволить машине стать по-настоящему интеллектуальной. Её решения складываются не только из ранее изученных ситуаций, но и могут порождать другие нестандартные или нелогичные решения, или ответы, как это может представить человек. Человек выстраивает свои логические цепочки основываясь не только на полученном ранее опыте, но и на основе своих догадок и предположений, а также на основе неявных связей между вещами и явлениями.

1. Логистическая регрессия и линейный дискриминант показали среднестатистический результат – AUC составляет 0,51 и 0,53, данные методы не предоставляют хорошее обучение модели вследствие не подходящих данных для обучения, но они линейный дискриминант показывает неплохую чувствительность и специфичность следует подбирать более лучшие параметры выборки данных

2. Метод случайный лес и бустинг показали неплохой результат обучение AUC 0,63 и 0,66 соответственно. Это не идеальный результат, бустинг требует более тонкой настройки для увеличения точность обучения.