Санкт-Петербургский Государственный Университет
Математико-механический факультет

Кафедра системного программирования

Минаев Александр Сергеевич

Использование линейного дискриминантного анализа в задаче предсказания оттока абонентов оператора сотовой связи

Курсовая работа

Научный руководитель:
д. ф.-м. н., профессор Терехов А. Н.

Санкт-Петербург
2018

Содержание

Введение

В настоящее время в сфере телекоммуникаций существует проблема оттока абонентов по причине возрастания конкуренции на рынке. Постоянно создаются новые тарифные планы, улучшаются условия старых, вследствие чего клиенты предпочитают уходить к конкурентам, предлагающим более выгодные условия. Например, у операторов мобильной связи ежегодный отток абонентов может достигать отметки в 50%. По этой причине компании несут огромные убытки.

Привлечение новых абонентов требует больше денег и ресурсов, чем удержание старых. Таким образом, задача предсказания оттока абонентов по данным об их активности – актуальная проблема. Будучи информированной о высокой вероятности ухода абонента, компания может предпринять ряд действий для его удержания.

Задача предсказания оттока абонента – это задача бинарной классификации. А именно, необходимо для каждого абонента определить, уйдет ли он в ближайшее время или нет. Применяя современные методы машинного обучения, можно получить хорошую точность предсказания, что позволяет компаниям создавать целые системы, которые предсказывают отток абонентов, дообучаются в процессе использования, а также своевременно предупреждают о возможности ухода особенно востребованных клиентов.

В данной работе изучается возможность применения линейного дискриминантного анализа для решения задачи предсказания оттока абонентов. Полученные результаты сравниваются с результатами работы других моделей.

1. Постановка задачи

Целью данной работы является реализация классификатора на основе линейного дискриминантного анализа, предсказывающего отток абонентов. Для ее достижения были поставлены следующие задачи:

- Проанализировать существующие решения
- Провести предобработку данных
- Разработать модель
- Оптимизировать результаты

2. Терминология

- **Обучающая выборка** — множество объектов, для которых известно, к каким классам они относятся. Используются для обучения модели.
- **Тестовая выборка** — выборка, по которой оценивается качество модели.
- **Классификация** — раздел машинного обучения, посвященный задаче построения алгоритма, классифицирующего произвольный объект на основе информации, полученной из обучающей выборки.
- **Классификатор** — модель, решающая задачу классификации.
- **Ансамбль классификаторов** — составная модель машинного обучения

3. Оценка эффективности

3.1. Метрики

- **Positive (P)** - положительный класс (уходящие абоненты)
- **Negative (N)** — отрицательный класс (остающиеся абоненты)
- **True Positive (TP)** — верно определенные в Positive
- **False Positive (FP)** — неверно определенные в Positive
- **True Negative (TN)** — верно определенные в Negative
- **False Negative (FN)** — неверно определенные в Negative
- **Accuracy** — процент верных предсказаний

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** - точность, доля верных P из предсказанных как P

$$precision = \frac{TP}{TP + FP}$$

- **Recall, TPR** — полнота, доля верных P из всех настоящих P

$$recall = \frac{TP}{TP + FN}$$

- **FPR** — доля неверно предсказанных N из настоящих N

$$FPR = \frac{FP}{FP + TN}$$

- **ROC** кривая — кривая, показывающая отношение TP к FP

- **AUC** — площадь под ROC кривой, интерпретируется как вероятность того, что случайно взятый Positive объект имеет оценку принадлежности к Positive выше, чем случайно взятый Negative объект.

3.2. Кросс-валидация

Для оценки эффективности моделей и проверки на переобучение используется кросс-валидация. Задается некоторое множество разбиений уже обработанных данных на обучающую и контрольную выборки. Для каждого разбиения выполняется настройка модели на обучающей выборке и производится оценка эффективности алгоритма на контрольной с помощью заданных выше метрик. Данным методом легко выявляется эффект переобучения моделей. Кросс-валидация является стандартной методикой тестирования и сравнения алгоритмов машинного обучения.

4. Обзор существующих решений

Задача предсказания оттока абонентов - актуальная проблема, поэтому существует большое число работ на эту тему. Для рассмотрения были выбраны несколько актуальных работ:

Автор	Работа	Метод
Antonio Canale and Nicola Lunardon	Churn Prediction in Telecommunications Industry. A Study Based on Bagging Classifiers [1]	Логистическая регрессия, линейный дискриминантный анализ, решающее дерево, нейронная сеть и др.
Naveen Kumar Rai and Vikas Srivastava and Rahul Kumar	Churn Prediction Model Using Linear Discriminant Analysis (LDA) [2]	Линейный дискриминантный анализ
Sahar F. Sabbeh	Machine-Learning Techniques for Customer Retention: A Comparative Study [3]	Линейный дискриминантный анализ, машина опорных векторов, случайный лес, нейронная сеть и др.
М.Корыстов	Применение методов машинного обучения для предсказания поведения абонентов сотовой связи [4]	Логистическая регрессия, нейронные сети, решающие деревья, бустинг
А. Сулягина	Оптимизация предсказания оттока абонентов оператора сотовой связи [5]	Бустинг, решающие деревья, нейронная сеть, метод ближайших соседей

В работах [1], [3], [4], [5] исследуется множество методов в задаче предсказания оттока клиентов. При этом, в работах [1], [3] среди рассматриваемых методов присутствует линейный дискриминантный анализ. В [1] приводятся результаты работы каждого из алгоритмов, а также бэггинга на них. Примечательно, что линейный дискриминантный анализ показывает один из худших результатов среди множества алгоритмов в работах [1] и [3], однако бэггинг над ним уже занимает второе место среди результатов бэггинга на других алгоритмах в [1].

В работе [2] линейный дискримантный анализ показывает крайне плохой результат в 74% точности. Учитывая тот факт, что в среднем отток клиентов составляет 25%, модель без машинного обучения, для каждого клиента утверждающая, что он не уйдет, имела бы аналогичный процент точности. Вероятно, данные были плохо обработаны перед тем, как их подали на вход алгоритму, или были неправильно выбраны признаки.

В работах [2] и [3] единственной используемой метрикой является точность. В задаче предсказания оттока клиентов из-за очевидного неравенства классов эта характеристика малоинформативна, если модели не оцениваются также по другим метрикам. Модель может иметь высокую точность, но пропускать многих уходящих клиентов и хорошо предсказывать остающихся. Такая модель является очень плохой с практической точки зрения. Поэтому необходимо использовать слабо зависимые от соотношения классов метрики, например, AUC.

Работы [4] и [5] не используют линейный дискриминантный анализ, однако они были выполнены на данных клиентов того же оператора сотовой связи, что и данная работа. Поэтому, имеет смысл сравнить лучший результат работы их моделей с результатами, полученными в ходе текущей работы.

5. Данные

Для обучения моделей использовались реальные данные пользователей одного из крупнейших в России операторов сотовой связи. Были предоставлены актуальные данные 300000 клиентов. Для каждого абонента имелась информация о последних 4 месяцев его активности. Также для каждого клиента было известно, ушел ли он. Таким образом, в наличии были размеченнные данные достаточного размера для обучения моделей.

Данные были агрегированы по следующим категориям:

- Минуты и стоимость входящих вызовов
- Минуты и стоимость исходящих вызовов
- Количество и стоимость исходящих СМС
- Количество входящих СМС
- Количество трафика мобильного интернета и его стоимость
- Информация об обращениях клиента в службу поддержки
- Тарифный план клиента
- Личные данные

Категории, связанные с трафиком, звонками и СМС, были разбиты отдельно на операции внутри субъекта РФ, страны, в роуминге.

6. Реализация

Вся работа выполнялась с помощью языка Python. Он хорошо подходит для научных вычислений, в том числе и машинного обучения, благодаря простоте языка, большому набору библиотек для решения задач обработки данных и их анализа.

Были использованы следующие библиотеки:

- **Pandas** - для обработки данных оператора сотовой связи.
- **Scikit-learn** - для построения и настройки моделей, оценки их эффективности.
- **Numpy** - для научных вычислений.
- **Matplotlib** - для построения графиков.

7. Линейный дискриминантный анализ

7.1. Описание метода

Линейный дискриминантный анализ (далее lda) - метод статистики и машинного обучения, который применяется для поиска оптимальных линейных комбинаций признаков, разделяющих пространство объектов на необходимое число классов. Таким образом, в задаче бинарной классификации с помощью lda находится оптимальная гиперплоскость (Рис. 1).

Для нахождения данной гиперплоскости решается вспомогательная задача. Требуется найти найти ось, проекция на которую максимизирует отношение общей дисперсии выборки к сумме дисперсий внутри отдельных классов. Результатом работы алгоритма будет вектор. Этот вектор является нормалью к искомой оптимальной гиперплоскости.

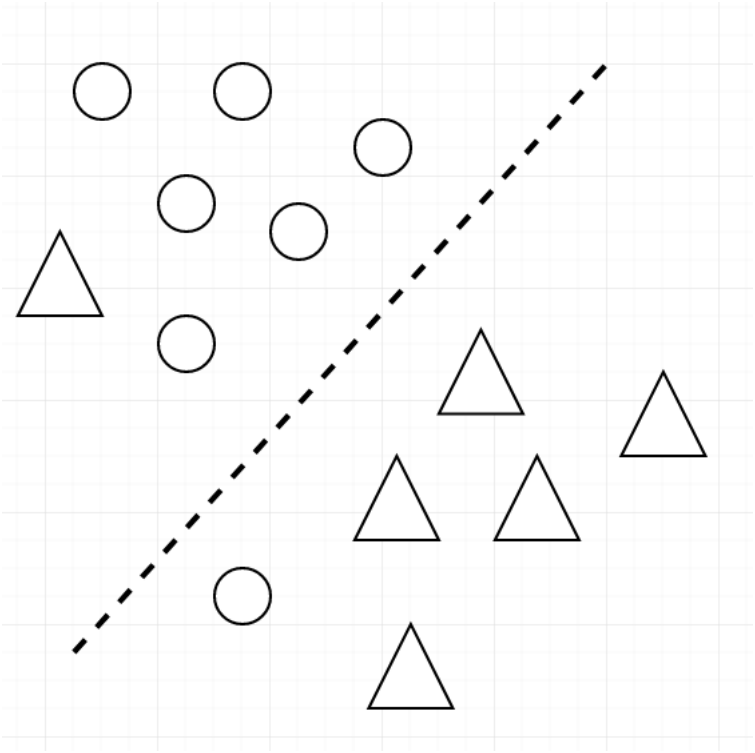

Рис. 1

7.2. Результаты

Оценка эффективности линейного дискриминантного анализа показала следующие результаты:

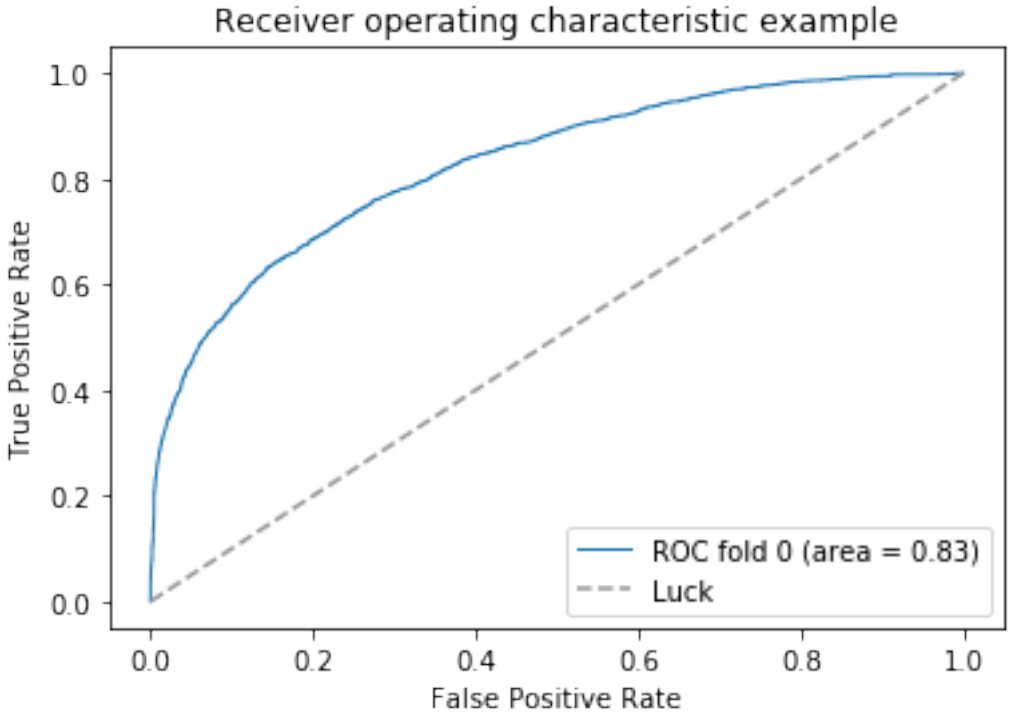

Рис. 2

AUC	Precision	Recall
0.83	0.54	0.58

Для оценки модели использовалась кросс-валидация с разбивкой данных на 10 частей. Изменение внутренних параметров модели не повлияло на результат.

Для дальнейшей оптимизации было решено использовать ансамбль из LDA, а именно бэггинг, поскольку в работе [1] он показал очень хорошие результаты в сравнении с базовыми алгоритмами.

8. Бэггинг над LDA

8.1. Описание метода

Бэггинг - это метод ансамблирования в машинном обучении. Используется несколько моделей одного и того же алгоритма, обученных независимо на своей части данных каждый. При классификации произвольного объекта его данные подаются на вход каждой из моделей. Окончательный результат выбирается путем голосования.

Бэггинг на подпространствах - схожий метод ансаблирования, отличающийся от описанного выше тем, что модели обучаются независимо не на разных объектах, а на разных признаках. Иначе говоря, каждая модель знает про всех объектов, но только про свою часть их признаков.

8.2. Результаты

После реализации бэггинга были получены следующие результаты:

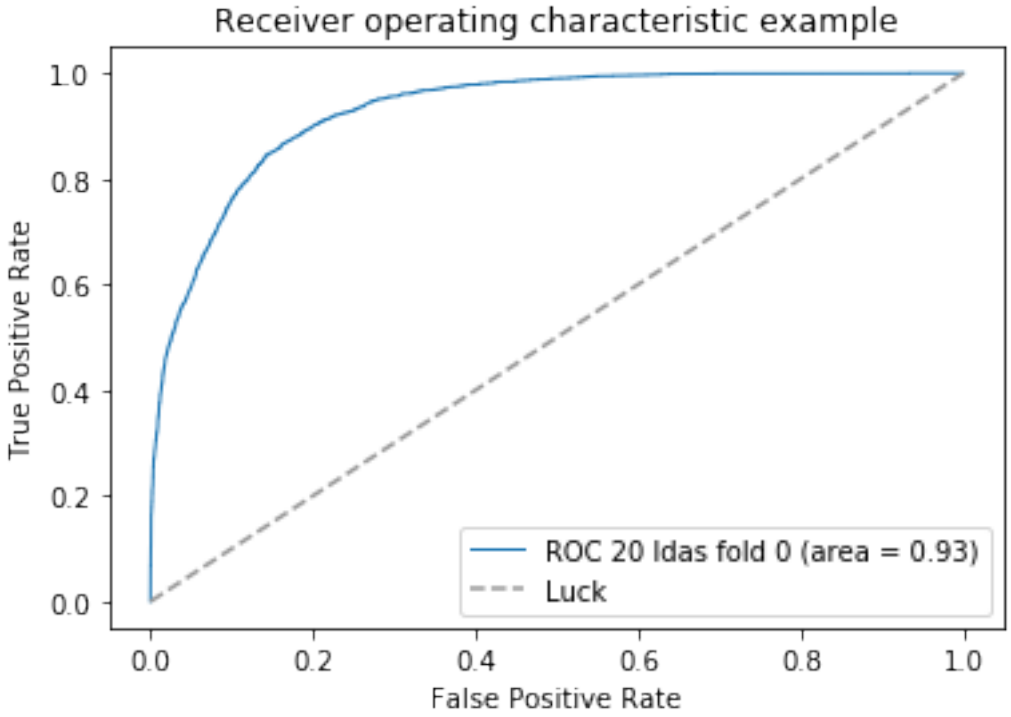

Рис. 3

AUC	Precision	Recall
0.93	0.67	0.75

Был выбран лучший результат, который достигался при голосовании 20 моделей. При большем числе уже наблюдалась деградация алгоритма вследствие недообученности. Можно заметить, что бэггинг действительно сильно улучшил результат, как и описывалось в работе [1].

Также был опробован бэггинг на подпространствах. Так как исходные данные были разбиты по месяцам, то признаки были разбиты следующим образом: за каждый месяц отвечал один LDA, остальные признаки были в отдельной модели. Учитывая, что месяцы неравнозначны (клиенту важнее что произошло в самом последнем месяце), был обучен линейный нейрон, который каждой модели присваивал вес в зависимости от важности ее признаков. Однако данный метод не показал улучшений результатов по сравнению с LDA.

Таким образом, сравнивая с лучшими результатами предыдущих лет полученными с помощью ансамбля из случайных лесов, градиентного бустинга и линейной регрессии в работе [5] (AUC - 0.92, Precision - 0.75, Recall - 0.72), можно заметить что были достигнуты те же результаты только с помощью бэггинга над LDA, и даже улучшены, если брать во внимание AUC и Recall, поскольку в первую очередь важно не пропустить уходящих абонентов.

Заключение

В рамках данной работы была изучена возможность применения линейного дискриминантного анализа и ансамблей основанных на нем для решения задачи оттока абонентов оператора сотовой связи. Были изучены существующие решения, обработаны полученные данные, реализованы несколько классификаторов основанных на линейном дискриминантном анализе, оценена их точность. Было проведено сравнение полученных результатов с результатами предыдущих лет.

Можно сделать вывод, что ансамбль построенный с помощью бэггинга над линейным дискриминантным анализом показывает хорошие результаты на предоставленных данных.

Список литературы

[1] Antonio Canale and Nicola Lunardon, *Churn Prediction in Telecommunications Industry. A Study Based on Bagging Classifiers*, Collegio Carlo Alberto, No.350, 2014

[2] Naveen Kumar Rai and Vikas Srivastava and Rahul Kumar, *Churn Prediction Model Using Linear Discriminant Analysis (LDA)*, IOSR Journal of Computer Engineering (IOSR-JCE), Volume 18, Issue 5, Ver. IV, 2016

[3] Sahar F. Sabbeh, *Machine-Learning Techniques for Customer Retention: A Comparative Study*, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 2, 2018

[4] Максим Корыстов, *Применение методов машинного обучения для предсказания поведения абонентов сотовой связи*, Дипломная работа, 2015

[5] Сулягина Анастасия *Оптимизация предсказания оттока абонентов оператора сотовой связи*, Курсовая работа, 2016