Применение машинного обучения к ранжированию инцидентов на Московской железной дороге

П.Ю. Бойко, Е.М. Быков, Е.И. Соколов, Д.А. Яроцкий

Аннотация. Московская железная дорога (МЖД) является крупной железнодорожной сетью, включающей в себя 8.8 тыс. км путей и 549 станций. МЖД оснащена несколькими десятками тысяч устройств автоматической регистрации отказов и предотказных состояний оборудования, сигналы которых обрабатываются операторами Центра управления содержанием инфраструктуры. Поток сигналов о возможных инцидентах создает большую нагрузку на операторов Центра. С целью оптимизации работы Центра была разработана основанная на машинном обучении система предварительного автоматического ранжирования инцидентов. Успешно внедренная предсказательная модель (ансамбль решающих деревьев) оценивает вероятность реального предотказного состояния по имеющимся признакам.

Ключевые слова: мониторинг инфраструктуры железной дороги, ранжирование инцидентов, машинное обучение, отбор признаков, ансамбль решающих деревьев.

Введение

Московская железная дорога (МЖД) является крупной железнодорожной сетью, включающей в себя 8.8 тыс. км путей и 549 станций. МЖД оснащена несколькими десятками тысяч устройств ЖАТ (железнодорожная автоматика и телемеханика), производящих автоматическую регистрацию инцидентов – отказов и предотказных состояний оборудования. Центр управления содержанием инфраструктуры Московской железной дороги (ЦУСИ МЖД) осуществляет непрерывный мониторинг инцидентов на всех участках железной дороги.

Инцидент представляет собой совокупность ситуаций, несущих признаки предотказного состояния (например, кратковременное превышение максимально допустимого напряжения оборудования). Наиболее распространенные типы ситуаций, приводящие к предотказам, показаны на Рис 1; всего выделяют около 600 различных типов ситуаций.

Ситуации автоматически объединяются в инциденты в базе данных ЦУСИ (Рис. 2). Текущие инциденты обрабатываются операторами ЦУСИ, которые выявляют их причины, классифицируют как относящиеся к одному из нескольких типов и принимают необходимые меры.

Выделяют следующие типы инцидентов (по причине возникновения): предотказное состояние, техническое обслуживание и ремонт, недостатки диагностики и технологическая ситуация. Предотказные состояния образуют наиболее важный тип инцидентов, требующий оперативного реагирования, однако его доля среди всех инцидентов сравнительно невелика (2-3%). При этом, благодаря достигнутой в последние годы высокой степени обеспечения систем МЖД средствами сбора данных, приходящий в ЦУСИ для обработки операторами поток инцидентов является крайне интенсивным: как правило, новые инциденты могут появляться с интервалом в несколько секунд; об-

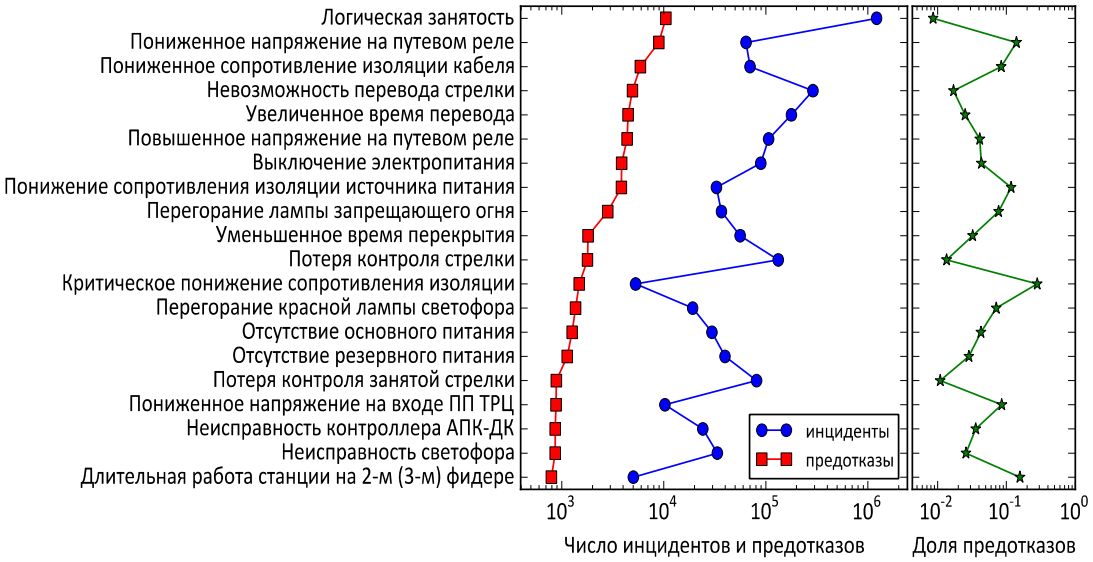

Рис. 1. Наиболее распространённые типы ситуаций, приводящих к предотказу

щее число инцидентов за 2015 год составило около 1.3 миллионов.

Таким образом, с точки зрения снижения нагрузки на операторов и повышения эффективности их работы крайне актуальной является задача предварительного автоматического ранжирования инцидентов по степени их важности. В статье описывается решение этой задачи, полученное нами с помощью методов машинного обучения и внедренное в систему принятия решений ЦУСИ МЖД.

Качество и себестоимость продукции и услуг зависят от большого количества параметров. Для управления процессом производства и обслуживания компании традиционно разрабатывают детализированные программы и спецификации. Регламентные подходы не всегда позволяют достичь максимальной точности принятия решений, приводя к дополнительным издержкам. Машинное обучение на основе имеющихся исторических данных все чаще используется исследователями для решения широкого класса индустриальных задач предсказательного обслуживания. В работе [1] случайный лес, градиентный бустинг и глубокая нейронная сеть используются для предсказания потребления топлива флотом авиакомпании. В [2, 3] авторы рассматривают задачу обнаружения сбоев в технологических процессах на производственной линии при помощи

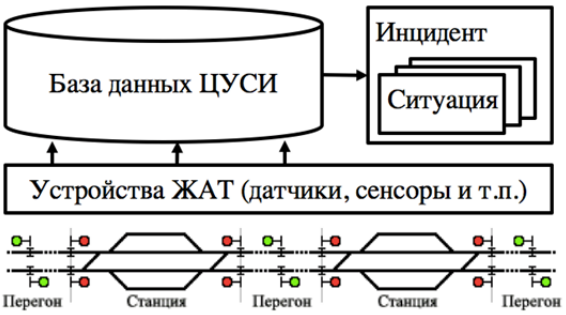

Рис. 2. Последовательность формирования инцидентов: архитектура системы мониторинга

моделей на основе логистической регрессии и случайного леса. В работе [4] авторы предлагают оптимизировать расход ферросплавов и добавочных материалов при производстве стали, в [5, 6] схожие методы используются для предсказания различных свойств химических составов. В большинстве случаев авторы фокусируются на сравнении алгоритмов и уделяют недостаточное внимание процессу порождения составных признаков и исследованию их индивидуальной эффективности при включении в модель. В данной статье мы отдельно описываем этот этап и предлагаем способ адаптивного добавления признаков к решающему правилу, что упрощает реализацию и внедрение указанных методов на практике.

1. Постановка задачи и методика решения

1.1. Статистический характер задачи

Отметим, что важность инцидента можно считать определяемой двумя факторами: вероятностью того, что данный инцидент является реальным предотказным состоянием, и последствиями (в частности, экономическим эффектом) данного отказа. Эти два фактора имеют совершенно разный характер и в значительной степени независимы друг от друга. В то время как вероятность предотказного состояния является предметом статистического анализа на основе лишь сохраненных оперативных данных об инцидентах, оценка последствий отказа предполагает привлечение дополнительных стоимостных моделей отказов и других соображений. В работе мы акцентируем внимание лишь на первом, вероятностном факторе важности; второй фактор учитывается ЦУСИ МЖД отдельно, с помощью нескольких альтернативных моделей. С учетом сделанного замечания цель данной работы можно сформулировать как построение прогнозной модели, которая предсказывает вероятность предотказного состояния для данного инцидента на основе автоматически сформированных данных об этом инциденте.

1.2. Исходные данные

Каждый в базе данных ЦУСИ представляет собой набор единичных событий, несущих признаки предотказного состояния (например, повышение напряжения, понижение сопротивления и т.п.) Рис. 2. Ситуации автоматически объединяются в инциденты на этапе предварительной обработки сигналов датчиков. В простейшем случае инцидент содержит несколько последовательных однотипных ситуаций (например, несколько случаев кратковременного повышения напряжения на одном и том же приборе), но может содержать и разнотипные ситуации, затрагивающие разное оборудование (например, увеличенное время перевода стрелки и некорректные параметры тока перевода). Количество ситуаций в инциденте может варьироваться от одного до нескольких сотен и может увеличиваться со временем, пока инцидент не рассмотрен и не закрыт оператором ЦУСИ. В общей сложности база данных ЦУСИ содержит около 5 млн инцидентов, включающих 100 млн. ситуаций.

Описание каждой ситуации содержит ее тип, время начала и окончания, ID места (станции или перегона), ID объекта измерения (прибора) и некоторые другие, менее существенные элементы. Большинство свойств (кроме времени) являются категориальными, причем количество их значений может быть достаточно большим: данные охватывают несколько сот типов ситуаций и мест и около 65000 приборов.

При закрытии инцидента оператором ставится пометка о выявленном типе инцидента, в частности являлся ли он предотказом.

Таким образом, исторические данные содержат достаточно богатую и хорошо структурированную информацию для построения сложных прогностических моделей вероятности предотказа.

1.3. Простейшая модель ранжирования

В качестве примера простой предсказательной модели можно привести прогноз вероятности предотказного состояния исключительно на основе типа первой ситуации инцидента, а именно как долю предотказных инцидентов среди всех наблюдавшихся ранее инцидентов с первой ситуацией данного типа. Это доля сильно зависит от типа ситуации, например, из Рис. 1 видно, что для ситуации «Критическое понижение сопротивления изоляции» она значительно выше, чем для «Потери контроля занятой стрелки». Поскольку различных типов ситуаций несколько сотен и операторы, очевидно, не в состоянии помнить характеристики каждой из них, даже эта простейшая автоматическая модель ранжирования оказывается практически полезной. В дальнейшем мы будем называть эту модель “референсной”.

1.4. Машинное обучение

Машинное обучение (МО) на основе имеющихся исторических данных об инцидентах позволяет строить гораздо более точные предсказательные модели. Стандартная практика машинного обучения [5, 6, 8] предполагает построение модели в два этапа:

1. извлечение признаков и формирование обучающей матрицы;

2. применение к полученной обучающей матрице некоторого общего МО-алгоритма.

В ходе первого этапа информация о каждом инциденте приводится к унифицированному виду числового вектора фиксированной размерности, компоненты которого (признаки) отражают потенциально существенные для прогноза характеристики инцидента. Этот этап описан в разделе 2.5.

В ходе второго этапа к полученной обучающей матрице обычно применяется один из стандартных МО-алгоритмов: логистическая регрессия, нейронные сети, ансамбли решающих деревьев и т.п. Нами использовался алгоритм XGBoost построения ансамбля решающих деревьев с помощью градиентного бустинга [9]. Этот этап описан в разделе 2.6.

1.5. Извлечение признаков

Извлечение признаков являлось наиболее творческой и трудоемкой частью решения задачи. Оно было связано со следующими трудностями.

• В общем случае, признаки необходимо агрегировать из всех ситуаций данного инцидента, учитывая, что ситуаций может быть произвольное число. Нами были рассмотрены различные стратегии агрегации, например: ограничиться первой или последней ситуацией в инциденте; в случае числовых признаков взять максимум, минимум или среднее значение по всем ситуациям; в случае категориальных признаков отметить все категории, встреченные в ситуациях инцидентах или сосчитать количество различных встреченных категорий. Конечно, некоторые признаки естественным образом связаны с инцидентом в целом (например, полная продолжительность или место инцидента).

• Естественным способом преобразования категориального признака с N возможными значениями в числовую форму является его кодирование в виде вектора длины N с единственным ненулевым элементом (one-hot-encoding). Ввиду того, что в рассматриваемой задаче N достигает нескольких сотен или даже тысяч, обучающие матрицы реализовывались нами в виде разреженных матриц.

С учетом этих обстоятельств, мы рассмотрели несколько десятков различных признаков, описывающих пространственно-временные и прочие характеристики инцидентов.

Важную роль при создании и отборе признаков играла оценка их значимости, которая осуществлялась нами с помощью двух методов.

• Во-первых, мы оценивали важность признака по общему числу соответствующих ветвлений в итоговой предсказательной модели – лесе решающих деревьев. Для “сильно ветвящихся” признаков мы производили попытку самостоятельно разбить признак на несколько вспомогательных.

• Во-вторых, мы реализовали жадный переборный алгоритм последовательного добавления в модель новых признаков, дающих наибольших прирост точности. Этот способ требует многократного обучения модели на различных наборах признаков и поэтому сравнительно долог, однако он позволяет понять, несут ли новые признаки какую-то существенную новую информацию по отношению к уже имеющимся, и какой минимальный набор признаков обеспечивает приемлемую точность модели. Мы подробно описываем результаты этого исследования в разделе 3.2.

Отметим, что помимо признаков, извлекаемых из базы инцидентов, нами была сделана попытка сформировать дополнительные признаки на основе метеоданных, поскольку погодные условия, очевидно, должны оказывать сильное влияние на возникновение некоторых предотказных состояний. Мы действительно обнаружили наличие корреляций между погодными признаками и количеством инцидентов, однако добавление этих признаков в нашу модель не дало заметного улучшения. Иными словами, информация о погоде позволяет улучшить прогноз возникновения инцидента при отсутствии иной информации, но не позволяет заметно улучшить прогноз предотказа при наличии уже имеющейся в базе данных информации об инциденте.

1.6. Ансамбль решающих деревьев

Предлагаемая предсказательная модель представляется в виде суммы ансамбля бинар-

ных решающих деревьев. В ходе обучения регрессионные решающие деревья последовательно добавляются в решающую композицию. Каждое решающее дерево имеет J листовых вершин, соответствующие J непересекающимся областям, на которые по признакам разбивается пространство инцидентов. Целевая функция, используемая при обучении модели, состоит из двух компонентов – стандартной функции потерь L на обучающем множестве и регуляризующей функции Ω, контролирующей сложность итоговой модели. Реализация предсказательной модели строилась нами с помощью библиотеки XGboost [9]. XGboost содержит эффективную реализацию градиентного бустинга [10], поддерживающую большие выборки и специальный формат хранения данных для операций с разреженными матрицами [11]. За последнее время эта открытая библиотека неоднократно использовалась исследователями для решения индустриальных задач, таких как предсказание сбоев технологических процессов [2, 3] и различных свойств химических составов [5, 6].

Заметим, что референсная модель из раздела 2.3 также естественным образом реализуется в виде ансамбля решающих деревьев, построенных по одному категориальному признаку – типу первого инцидента. А именно, она состоит из тривиальных деревьев глубины 1 (“решающих пней”), по одному на каждую из компонент бинарного вектора, кодирующего этот категориальный признак. Таким образом, основная модель может считаться естественным обобщением референсной модели на случай произвольного набора признаков, числа деревьев и их глубины.

Основная модель была обучена по 3 годам исторических данных (около 5.3 млн инцидентов) и включала в себя 4000 деревьев глубины 10. Подбор параметров модели осуществлялся стандартным образом с помощью тестирования на контрольной части обучающей выборки.

2. Анализ модели

2.1. Тестовое ранжирование

В соответствии со стандартной практикой машинного обучения, тестирование построенной модели проводилось на тестовой выборке, изолированной от обучающей выборки, причем обучающие инциденты предшествовали по времени тестовым.

На Рис. 3 приведены результаты ранжирования 10000 последовательных инцидентов. Крупные точки показывают инциденты, связанные с предотказами, а мелкие – инциденты, связанные с ложными срабатываниями системы. Хорошо видно, что основная масса предотказов имеет высокую вероятность и визуально отделена по вероятности от основной массы ложных тревог, имеющей меньшую вероятность. Это свидетельствует о пригодности

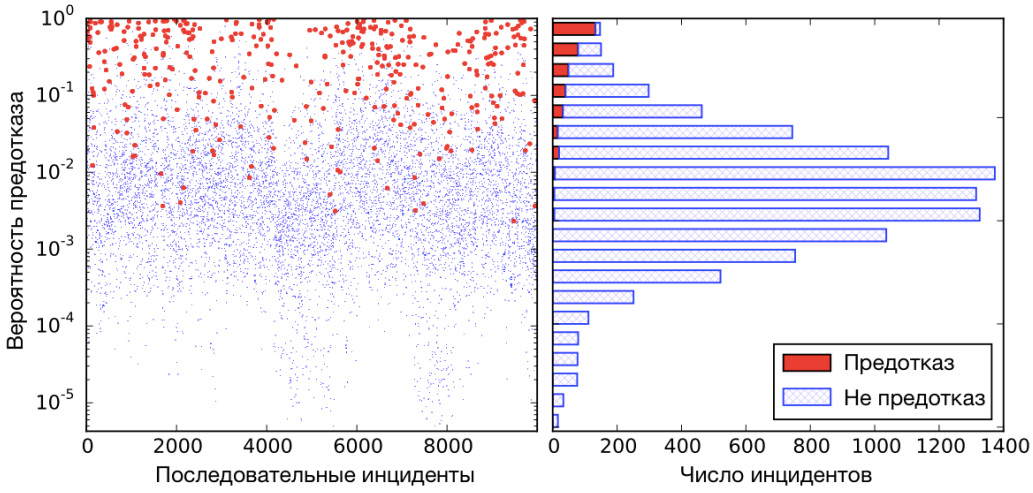

Рис. 3. Результаты ранжирования 10000 последовательных инцидентов

Предотказы показаны крупными точками, ложные тревоги – мелкими. В правой части рисунка показаны соответствующие гистограммы распределения инцидентов по вероятности

использования предсказательной модели в качестве ранжирующей функции.

Из рисунка очевидна некоторая неравномерность и периодичность распределения вероятности предотказа в зависимости от номера инцидента. Этот эффект объясняется суточным циклом (показанные 10000 инцидентов отвечают временному интервалу продолжительностью около 3 дней). В дневное время значительное число инцидентов связано с техническим обслуживанием и ремонтом; модель присваивает низкую вероятность предотказа таким инцидентам.

2.2. Количественные оценки эффективности ранжирования

Для количественной оценки эффективности ранжирования мы используем кривую ошибок, определяемую следующим образом. Предположим, что операторы успевают обрабатывать лишь некоторую долю всех инцидентов; рассмотрим соответствующую переменную ДОИ (Доля Обработанных Инцидентов) со значениями в интервале $[0,1]$. Будем считать, что операторы в первую очередь обрабатывают те инциденты, для которых модель предсказывает наиболее высокую вероятность отказа. Пусть величина ДОО (Доля Охваченных Отказов), также со значениями в интервале $[0,1]$, обозначает долю обрабатываемых при этом предотказных инцидентов среди всех предотказных инцидентов. Кривая ошибок тогда представляет собой график зависимости ДОО от ДОИ.

На Рис. 4 показаны кривые ошибок для основной и референсной моделей ранжирования. Чем выше лежит кривая, тем лучше. Максимально возможное положение кривой соответствует линейной функции $\text{ДОО} = \frac{\text{ДОИ}}{\alpha}$, где $\alpha \approx 0.024$ – доля предотказов среди всех инцидентов. Если оператор выбирает инцидент наугад (без рассмотрения ранга), то кривая ошибок является диагональю квадрата.

В качестве основной количественной характеристики точности мы рассматриваем AUC (Area Under Curve) – площадь под кривой ошибок. Ее значения для обеих моделей приведены в Табл. 1. Кроме того, мы приводим значения ДОИ95% и ДОИ99%, которые определяются

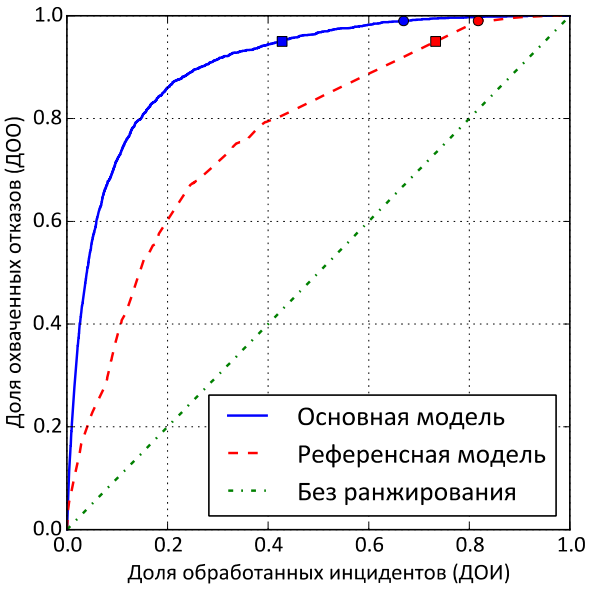

Рис. 4. Кривые ошибок и связанные с ними характеристики для основной и референсной моделей

Выделенные точки отвечают ДОО 0.95 и 0.99

Табл. 1. Сводка результатов для основной и референсной моделей

Модель	AUC	ДОО 0.95	ДОО 0.99
Основная	0.904	0.427	0.669
Референсная	0.768	0.733	0.818

как те значение ДОИ, при которых ДОО составляет, соответственно, 0.95 и 0.99. Заметим, в частности, что если операторы обрабатывают лишь половину инцидентов, а именно те, которые имеют основной ранг выше медианного значения, то при этом пропускается менее 5% предотказных состояний.

2.3. Анализ эффективности признаков

На Рис. 5 показаны результаты эксперимента по последовательному адаптивному добавлению в модель новых признаков. На каждом шаге для каждого из еще не входящих в модель признаков совершается пересчет точности модели с этим дополнительным признаком; тот, для которого наблюдается наибольшее приращение точности, включается в модель. В общей сложности рассматривается 34 признака, процедура начинается с признака типа первой ситуации (как наиболее информативного).

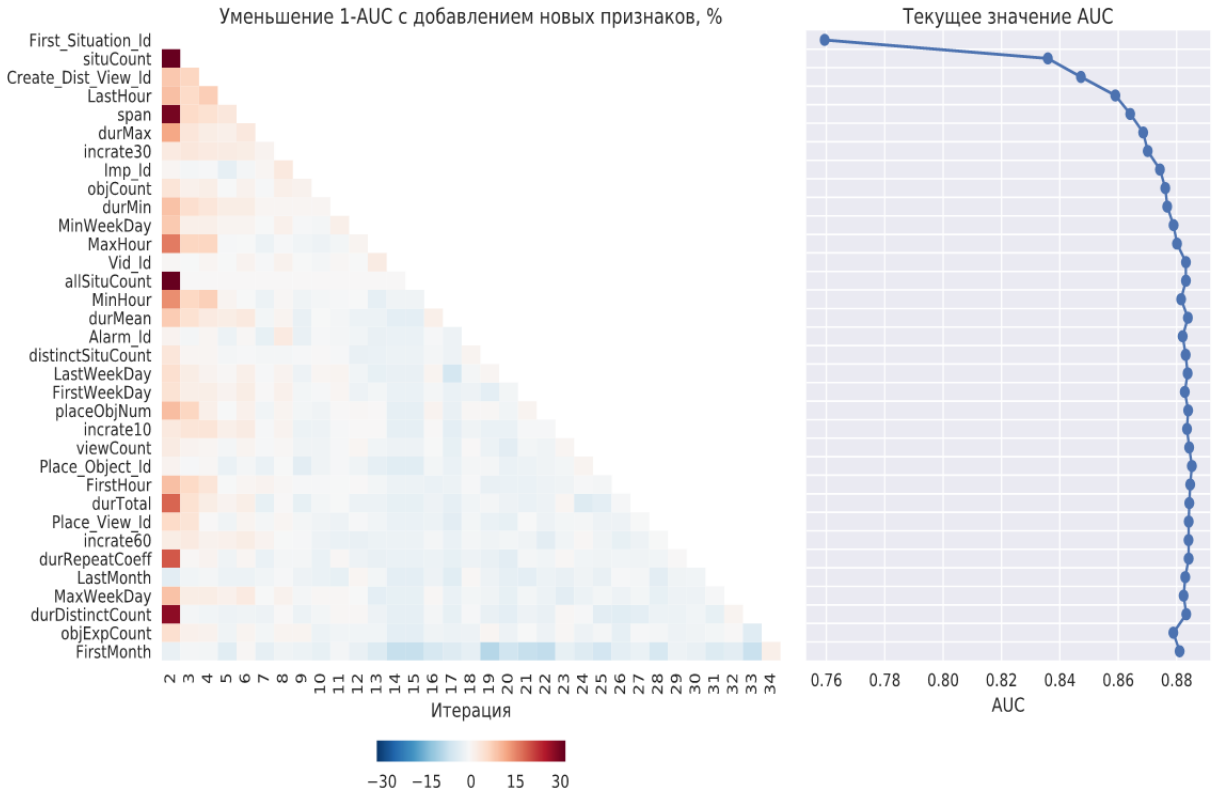

Рис. 5. Анализ важности признаков с помощью их последовательного жадного добавления

На каждой итерации к текущему набору признаков, начиная с First_Situation_id, добавляется новый признак, при котором точность ранжирования увеличивается сильнее всего. (a) На диаграмме слева показаны относительные изменения величины 1-AUC при добавлении каждого из признаков, не входящих в текущий набор. График справа показывает динамику точности предсказательной модели, построенной по текущему набору признаков. (b) Несколько первых признаков, до насыщения модели

В левой части Рис. 5 a показаны относительные изменения характеризующей ошибку величины $1 - \mathrm{AUC}$ для каждой итерации (горизонтальная ось) и каждого признака–кандидата (вертикальная ось). Признаки упорядочены post factum по очередности включения в модель, так что диаграмма имеет треугольный вид. В правой части рисунка показаны значения AUC соответствующей модели для каждой итерации. Для ускорения процедуры, модели в этом эксперименте строились с уменьшенным числом деревьев, поэтому их значения AUC ниже значения, указанного для основной модели на Рис. 4.

В результате данного эксперимента мы видим, что модель насыщается после включения в нее 10 – 15 признаков, и ее точность перестает существенно меняться, если не считать небольшого эффекта переобучения, наблюдаемого в конце эксперимента. Первые несколько признаков являются одновременно достаточно информативными и разнообразными; их описания приведены в Табл. 2. На левой части Рис. 5 a хорошо видно, что полезность признаков падает с итерациями, причем падение является резким в моменты, когда в набор включается признак, родственный рассматриваемому (например, *span* и *situCount*). Несмотря на это, мы видим, что полезно иметь, например, много разных признаков, характеризующих время ситуаций.

3. Опыт внедрения

Построенная модель ранжирования инцидентов была интегрирована в виде нового элемента, АПК «САРИ» (Система автоматического ранжирования инцидентов), в комплекс диспетчерского контроля ЦУСИ РЖД. Схема интеграции и взаимодействующие модули представлены на Рис. 6.

Табл. 2. Наиболее эффективные признаки, в порядке включения в модель

Признак	Описание	Тип / кол-во значений
First_Situation_id	Тип первой ситуации в инциденте	Категориальный, 600
situCount	Количество ситуаций в инциденте	Количественный
Create_Dist_View_id	Дистанция (место) инцидента	Категориальный, 22
LastHour	Час последней ситуации в инциденте	Количественный
span	Продолжительность инцидента от начала первой до начала последней ситуации	Количественный
durMax	Максимальная продолжительность ситуации в инциденте	Количественный
incrate30	Количество инцидентов, зарегистрированных на объекте в пределах 30 мин до регистрации инцидента	Количественный
imp_id	Степень важности инцидента	Категориальный, 4
objCount	Количество объектов, затронутых инцидентом	Количественный
durMin	Минимальная продолжительность ситуации в инциденте	Количественный
MinWeekDay	Минимальное значение дня недели среди ситуаций инцидента	Категориальный, 7

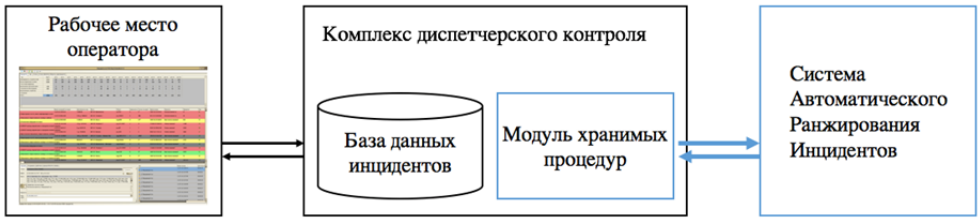

Рис. 6. Схема интеграции САРИ и АПК ДК

Модуль ранжирования инцидентов выполняется на изолированном сервере. Он синхронно опрашивает модуль хранимых процедур в базе данных инцидентов на предмет обновления полей инцидентов и при необходимости инициирует пересчет при получении новых инцидентов и ситуаций. Важно отметить, что не все поля данных могут быть получены в оперативном режиме.

При получении новых данных модуль осуществляет ранжирование инцидентов по набору исходных признаков и выдает прогнозное значение. Это значение далее отображается в виде отдельного столбца в интерфейсе инженера службы мониторинга с возможностью сортировки (Рис. 7). Предсказательная модель выдает обновленное ранжирующее значение за 1.2 миллисекунды. Для удобства восприятия операторами, ранги инцидентов выдаются в виде целых чисел из интервала $[0{,}100]$ и соответствуют интервалу вероятностей $[10^{-6}, 1]$ на логарифмической шкале.

За время подконтрольной эксплуатации системы (май–июнь 2016 года, 1 месяц) в ней было зарегистрировано 581 032 ситуаций, объединенных в 92 339 инцидентов. Среди них предотказами было признано 2 187 инцидентов. Точность модели в тестовой эксплуатации оказалась очень близка к предварительной оценке на тестовой выборке (AUC 0.901 и 0.904, соответственно).

Заключение

Нами была разработана и внедрена основанная на машинном обучении модель автоматического ранжирования инцидентов, регистрируемых устройствами ЖАТ на Московской железной дороге. Модель позволяет инженерам

Рис. 7. Отображение результатов ранжирования в пользовательском интерфейсе инженера службы мониторинга

Центра управления инфраструктурой оперативно оценивать вероятности отказов и предотказных состояний и сокращает время реакции и устранения неполадок. Нами продемонстрирована высокая точность ранжирования с помощью построенной модели. Дополнительно мы также продемонстрировали метрику эффективности и способ адаптивного добавления составных признаков. Нам представляется полезным и перспективным более широкое внедрение предложенной нами технологии в системах железнодорожного мониторинга.

Отметим два возможных направления развития данного проекта. Во-первых, как обсуждалось в разделе 2.1, разработанная нами модель является чисто статистической и не учитывает различия в важности разных отказов и предотказных состояний. Было бы полезно дополнить нашу модель достаточно точной программной моделью потерь как функции отказа. Математически это можно выразить следующим образом. Имеющаяся модель оценивает условную вероятность отказа или предотказного состояния при наличии инцидента с определенными характеристиками: $\mathbb{P}(\text{предотказ}|\text{инцидент})$. Практически более полезной была бы оценка условного математического ожидания потерь (например, в денежном выражении) при наличии данного инцидента $\mathbb{E}(\text{потери}|\text{отказ})$. Если считать, что потери происходят только при предотказе, то $\mathbb{E}(\text{потери}|\text{отказ}) = \mathbb{E}(\text{потери}|\text{предотказ}) \cdot \mathbb{P}(\text{предотказ}|\text{инцидент})$, то есть для оценки $\$\,\mathbb{E}(\text{потери}|\text{инцидент})$ нам необходимо перемножить разработанную нами оценку вероятности предотказа на ожидаемые потери в случае предотказа, $\mathbb{E}(\text{потери}|\text{предотказ})$. Оценка этой последней величины может быть либо задана явно в соответствии с существующими регламентами, либо, при наличии хорошо структурированных и достаточно полных исторических данных о потерях, получена посредством их анализа с помощью методов, аналогичных применявшимся в настоящей работе.

Другим возможным направлением развития проекта является более глубокий анализ инцидентов, возможно, с привлечением более детальной информации о предотказных ситуациях. Имеющаяся в основной исторической базе данных информация об инцидентах часто довольно скудна, например, если инцидент состоит лишь из одной ситуации, описывающей повышение напряжения, то вся доступная информация об этом событии ограничивается указанием соответствующего временного интервала. В то же время можно ожидать, что дополнительная информация, скажем, о максимальном значении напряжения, скорости его изменения и т.п., позволила бы уточнить прогноз предотказного состояния.

Литература

1. Horituchi, Yuji, Baba, Yukino, Kashima, Hisashi, Suzuki, Masahito, Kayahara, Hiroki, & Maeno, Jun. 2017. Predicting Fuel Consumption and Flight Delays for Low-Cost Airlines. In: Twenty-Ninth IAAI Conference.
2. Hebert, Jeff. 2016. Predicting rare failure events using classification trees on large scale manufacturing data with complex interactions. Pages 2024–2028 of: 2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016.
3. Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. 2001. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.
4. YDF's Recommender System to Decrease Steelmaking Costs at Magnitogorsk Iron and Steel Works, https://yandexdatafactory.com/ru/company/press/magnitogorsk-iron-steel-works-save-4-million-annually-data-analytics/ Accessed: 2017-02-25.
5. Mitchell, Thomas M. 1997. Machine Learning. 1 edn. New York, NY, USA: McGraw-Hill, Inc. Mustapha, Ismail Babajide, & Saeed, Faisal. 2016. Bioactive molecule prediction using extreme gradient boosting. Molecules, 21(8), 983.
6. Sheridan, Robert P, Wang, Wei Min, Liaw, Andy, Ma, Junshui, & Gifford, Eric M. 2016. Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. Journal of Chemical Information and Modeling, 56(12), 2353–2360.
7. Kaggle. 2016. The Bosch Production Line Performance competition. https://www.kaggle.com/c/ bosch-production-line-performance. Accessed: 2017-02-25.
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten- hofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
9. Chen, Tianqi, & Guestrin, Carlos. 2016. Xgboost: A scalable tree boosting system. arXiv preprint arXiv:1603.02754.
10. Friedman, Jerome H. 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232.
11. Cerqueira V., Pinto F., Sá C., Soares C. (2016) Combining Boosted Trees with Metafeature Engineering for Predictive Maintenance. In: Boström H., Knobbe A., Soares C., Papapetrou P. (eds) Advances in Intelligent Data Analysis XV. IDA 2016. Lecture Notes in Computer Science, vol 9897. Springer, Cham.

Бойко Павел Юрьевич. Старший научный сотрудник Сколковского института науки и технологий, г. Москва. Окончил Московский физико-технический институт в 2005 году. Кандидат физико-технических наук. Количество печатных работ: 30. Область научных интересов: прикладная математика, распознавание образов, информационные технологии. E-mail: p.boyko@skoltech.ru

Быков Евгений Михайлович. Научный сотрудник, аспирант ИППИ им. А.А. Харкевича. Окончил Московский физико-технический институт в 2013 году. Количество печатных работ: 7. Область научных интересов: прикладная математика, информационные технологии. E-mail: e.bykov@skoltech.ru

Евгений Иванович Соколов. Заместитель начальника Технического центра автоматики и телемеханики ОАО «РЖД». Глава совместной рабочей группы, занимающейся разработкой системы автоматического ранжирования инцидентов. Область научных интересов: прикладная математика, теория управления. E-mail: e.sokolov@msk.rzd.ru

Дмитрий Александрович Яроцкий. Старший научный сотрудник Сколковского института науки и технологий, г. Москва. Окончил МГУ им. Ломоносова в 1998 году. Доктор физико-технических наук. Количество печатных работ: 15. Область научных интересов: прикладная математика, анализом данных. E-mail: d.yarotsky@skoltech.ru

Application of Machine Learning to Incident Ranking at Moscow Railway

P.Y. Boyko, E.M. Bikov, E.I. Sokolov, D.A. Yarotsky

Abstract. Moscow Railway, a large railway network including 8800 kilometers of track and 549 stations, is equipped with tens of thousands of devices for automatic registration of system failures. Alerts produced by these devices are processed by operators of the Infrastructure Management Center. The alert flow is very intense and creates a significant stress on the operators while about 97% of the signals turn out to be false alarms. To optimize the operation of the Center we have used machine learning to develop an advanced automated incident ranking model that estimates the probability of an actual failure from multiple features of the registered incident. The model was trained as an ensemble of decision trees by the algorithm XGBoost using a database of 5 million historical incidents. The model has been integrated into the software infrastructure of the Center and is used in the daily work of operators.

Keywords: railroad monitoring, incident ranking, machine learning, feature engineering, ensemble of decision trees, XGBoost.

References

1. Horituchi, Yuji, Baba, Yukino, Kashima, Hisashi, Suzuki, Masahito, Kayahara, Hiroki, & Maeno, Jun. 2017. Predicting Fuel Consumption and Flight Delays for Low-Cost Airlines. In: Twenty-Ninth IAAI Conference.
2. Hebert, Jeff. 2016. Predicting rare failure events using classification trees on large scale manufacturing data with complex interactions. Pages 2024–2028 of: 2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016.
3. Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. 2001. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.
4. YDF's Recommender System to Decrease Steelmaking Costs at Magnitogorsk Iron and Steel Works, https://yandexdatafactory.com/ru/company/press/magnitogorsk-iron-steel-works-save-4-million-annually-data-analytics/ Accessed: 2017-02-25.
5. Mitchell, Thomas M. 1997. Machine Learning. 1 edn. New York, NY, USA: McGraw-Hill, Inc. Mustapha, Ismail Babajide, & Saeed, Faisal. 2016. Bioactive molecule prediction using extreme gradient boosting. Molecules, 21(8), 983.
6. Sheridan, Robert P, Wang, Wei Min, Liaw, Andy, Ma, Junshui, & Gifford, Eric M. 2016. Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. Journal of Chemical Information and Modeling, 56(12), 2353–2360.
7. Kaggle. 2016. The Bosch Production Line Performance competition. https://www.kaggle.com/c/ bosch-production-line-performance. Accessed: 2017-02-25.
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten- hofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
9. Chen, Tianqi, & Guestrin, Carlos. 2016. Xgboost: A scalable tree boosting system. arXiv preprint arXiv:1603.02754.
10. Friedman, Jerome H. 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232.
11. Cerqueira V., Pinto F., Sá C., Soares C. (2016) Combining Boosted Trees with Metafeature Engineering for Predictive Maintenance. In: Boström H., Knobbe A., Soares C., Papapetrou P. (eds) Advances in Intelligent Data Analysis XV. IDA 2016. Lecture Notes in Computer Science, vol 9897. Springer, Cham.

Dmitry Yarotsky graduated from the Department of Mechanics and Mathematics of Moscow State University in 1998, where he also obtained his Candidate of Sciences degree in 2002. Doctor of Sciences. degree from IITP. Scientific interests: applied mathematics, data analysis, and optimization.

Evgeny Sokolov is Deputy Head of the Technical Center for Automation and Remote control at OJSC "Russian Railways". He leaded member of a joint working group developing a system for automatic incidents ranking.

Pavel Boyko received his Candidate of Sciences degree in Theoretical Physics from Alikhanov Institute for Theoretical and Experimental Physics in 2008. Doctor of Sciences. Scientific interests: applied mathematics, data analysis, and optimization.

Evgeni Bikov graduated from the Radio Engineering Faculty of MIPT in 2013, PhD student. He later worked at the Institute for Information Transmission Problems (IITP) in wireless data laboratory. Scientific interests: applied mathematics, computer science.