**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**

**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ**

**Кафедра дискретной математики и алгоритмики**

ПРУДНИКОВ

Евгений Александрович

**ОТЧЕТ № 2 О РАБОТЕ В РАМКАХ МАГИСТЕРСКОЙ ДИССЕРТАЦИИ**

специальность 1-31 81 09 «Алгоритмы и системы обработки больших объемов информации»

Минск, 2017

**ОГЛАВЛЕНИЕ**

[ПОСТАНОВКА ЗАДАЧИ 3](#_Toc478377828)

[ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ ПРЕДУПРЕЖДЕНИЯ ОШИБОК НА ПРОИЗВОДСТВЕННЫХ ЛИНИЯX 4](#_Toc478377829)

[Физические 4](#_Toc478377830)

[Программные 5](#_Toc478377831)

[ОБЗОР МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ 6](#_Toc478377832)

[Машинное обучение 6](#_Toc478377833)

[Общая постановка задачи обучения по прецедентам 6](#_Toc478377834)

[Основные стандартные типы задач 7](#_Toc478377835)

[Выводы 10](#_Toc478377836)

[СРЕДСТВА РАЗРАБОТКИ 11](#_Toc478377837)

[АНАЛИЗ ДАННЫХ, ГЕНЕРИРУЕМЫХ ДАТЧИКАМИ ПРОИЗВОДСТВЕННОЙ ЛИНИИ 12](#_Toc478377838)

[Анализ 12](#_Toc478377839)

[Выбор метрики качества модели 14](#_Toc478377840)

[ПОСТРОЕНИЕ МОДЕЛИ ПРЕДУПРЖДЕНИЯ СБОЕВ НА ПРОИЗВОДСТВЕННОЙ ЛИНИИ 15](#_Toc478377841)

[Описание алгоритма 15](#_Toc478377842)

[ЗАКЛЮЧЕНИЕ 16](#_Toc478377843)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 17](#_Toc478377844)

[ПРИЛОЖЕНИЕ А 18](#_Toc478377845)

ПОСТАНОВКА ЗАДАЧИ

Производство и производственные линии стали неотъемлемой частью современного общества. Обеспечение строгого следования стандартам качества и безопасности является одной из главных задач предприятия. Современные производственные линии снабжены большим количеством различных датчиков, которые позволяют собирать информацию о производственном процессе. Обработка этой информации и разработка, на её основе, встраиваемой модели предупреждения сбоев поможет повысить эффективность производственной линии, а также предотвратить дополнительные финансовые издержки и, следовательно, удешевить продукт для конечного пользователя.

Целью магистерской диссертации является разработка моделей предупреждения сбоев на производственной линии с использованием методов машинного обучения. Исходя из указанной цели, можно выделить частные задачи, поставленные в магистерской диссертации:

* Обработка измерительных данных датчиков производственной линии
* Разработка моделей анализа обработанных данных для оценки вероятности возникновения сбоя на производственной линии

ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ ПРЕДУПРЕЖДЕНИЯ ОШИБОК НА ПРОИЗВОДСТВЕННЫХ ЛИНИЯX

В настоящее время современные производственные линии, используемые предприятиями, обычно оснащены множеством систем предупреждения ошибок и контроля качества продукции, а также системами помощи оператору.

Условно можно разделить эти системы на два класса: физические, то есть различные датчики сбора оперативной информации, и программные, то есть программы, которые обрабатывают информацию с датчиков и предоставляют её в удобном виде, а также программы помощи операторам линии.

Физические

К физическим системам обнаружения и предотвращения ошибок можно отнести различные датчики контроля движения детали по производственной линии, анализирующие расположение и траекторию перемещения материалов по производственной линии. Данные элементы бывают нескольких типов:

*Концевые датчики* - контролируют положение рабочих органов станка. Срабатывают, когда подвижный элемент оборудования достигает определенной точки. Чаще всего устанавливают такие элементы на осях роликовых платформ и конвейерных лент. При нарушении правильного положения ленты или расположенных на ней материалов концевик информирует оператора, что в работе оборудования произошел сбой.

*Оптические линейки* – набор оптоэлектронных датчиков, с помощью которых осуществляется измерение габаритов материалов, движущихся по производственной линии. Такие линейки помогают оценить точность манипуляций, проводимых с материалом на деревообрабатывающих и режущих станках. Информация, получаемая с помощью оптической линейки, в форме аналогового сигнала передается на УЦИ и визуализируется на дисплее в виде цифр.

*Ультразвуковые датчики* – осуществляют контроль и измерение расстояния от предмета до заданного объекта. Могут применяться для определения уровня прогиба полотна, скорости работы станка, промежутков между материалами на технологической ленте. Используются в отраслях деревообработки, машиностроения, производства пищевых продуктов и на предприятиях целлюлозно-бумажной промышленности.

*Лазерные датчики (дальномеры)* – оптические устройства для определения расстояния между объектами. Работа основана на принципе измерения времени пролета луча, диапазон действия – до нескольких десятков метров. Применяют дальномеры в различных областях промышленности.

*Индуктивные датчики* – это детекторы металла. Установив такие элементы на производственной линии, можно вести подсчет изготовленных металлических деталей, осуществлять распознавание и сортировку деталей различного типа, контролировать скорость подачи материалов с одного этапа производства на другой. Применяться такие датчики могут везде, где ведется работа с элементами из металла.

Программные

К программным средствам контроля качества изделий, а также помощи оператору можно отнести различное ПО для сбора (базы данных, ERP системы), анализа и отображения информации, поступающей с датчиков, установленных на производственной линии. Примерами могут служить такие компании как Videojet с ПО CLARiTY, Omron, Balluff. Также, и собственные системы производителей.

ОБЗОР МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

В настоящее время активно используются системы, построенные на основе алгоритмов машинного обучения. Такие системы сейчас используются во многих сферах науки и техники. Учитывая задачу, а также исходные данные мы можем обратиться к алгоритмам машинного обучения для её решения.

Машинное обучение

Машинное обучение (Machine Learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. Различают два типа обучения. Обучение по прецедентам, или индуктивное обучение, основано на выявлении общих закономерностей по частным эмпирическим данным. Дедуктивное обучение предполагает формализацию знаний экспертов и их перенос в компьютер в виде базы знаний. Дедуктивное обучение принято относить к области экспертных систем, поэтому термины машинное обучение и обучение по прецедентам можно считать синонимами.

Машинное обучение находится на стыке математической статистики, методов оптимизации и классических математических дисциплин, но имеет также и собственную специфику, связанную с проблемами вычислительной эффективности и переобучения. Многие методы индуктивного обучения разрабатывались как альтернатива классическим статистическим подходам. Многие методы тесно связаны с извлечением информации и интеллектуальным анализом данных (Data Mining).

Наиболее теоретические разделы машинного обучения объединены в отдельное направление, теорию вычислительного обучения (Computational Learning Theory, COLT).

Машинное обучение — не только математическая, но и практическая, инженерная дисциплина. Чистая теория, как правило, не приводит сразу к методам и алгоритмам, применимым на практике. Чтобы заставить их хорошо работать, приходится изобретать дополнительные эвристики, компенсирующие несоответствие сделанных в теории предположений условиям реальных задач. Практически ни одно исследование в машинном обучении не обходится без эксперимента на модельных или реальных данных, подтверждающего практическую работоспособность метода.

Общая постановка задачи обучения по прецедентам

Дано конечное множество прецедентов (объектов, ситуаций), по каждому из которых собраны (измерены) некоторые данные. Данные о прецеденте называют также его описанием. Совокупность всех имеющихся описаний прецедентов называется обучающей выборкой. Требуется по этим частным данным выявить общие зависимости, закономерности, взаимосвязи, присущие не только этой конкретной выборке, но вообще всем прецедентам, в том числе тем, которые ещё не наблюдались. Говорят, также о восстановлении зависимостей по эмпирическим данным — этот термин был введён в работах Вапника и Червоненкиса.

Наиболее распространённым способом описания прецедентов является признаковое описание. Фиксируется совокупность n показателей, измеряемых у всех прецедентов. Если все n показателей числовые, то признаковые описания представляют собой числовые векторы размерности n. Возможны и более сложные случаи, когда прецеденты описываются временными рядами или сигналами, изображениями, видеорядами, текстами, попарными отношениями сходства или интенсивности взаимодействия, и т. д.

Для решения задачи обучения по прецедентам в первую очередь фиксируется модель восстанавливаемой зависимости. Затем вводится функционал качества, значение которого показывает, насколько хорошо модель описывает наблюдаемые данные. Алгоритм обучения (learning algorithm) ищет такой набор параметров модели, при котором функционал качества на заданной обучающей выборке принимает оптимальное значение. Процесс настройки (fitting) модели по выборке данных в большинстве случаев сводится к применению численных методов оптимизации.

Основные стандартные типы задач

*Обучение с учителем (supervised learning)*— наиболее распространённый случай. Каждый прецедент представляет собой пару «объект, ответ». Требуется найти функциональную зависимость ответов от описаний объектов и построить алгоритм, принимающий на входе описание объекта и выдающий на выходе ответ. Функционал качества обычно определяется как средняя ошибка ответов, выданных алгоритмом, по всем объектам выборки.

* *Задача классификации (classification)* отличается тем, что множество допустимых ответов конечно. Их называют метками классов (class label). Класс — это множество всех объектов с данным значением метки.
* *Задача регрессии (regression)* отличается тем, что допустимым ответом является действительное число или числовой вектор.
* *Задача ранжирования (learning to rank)* отличается тем, что ответы надо получить сразу на множестве объектов, после чего отсортировать их по значениям ответов. Может сводиться к задачам классификации или регрессии. Часто применяется в информационном поиске и анализе текстов.
* *Задача прогнозирования (forecasting)* отличается тем, что объектами являются отрезки временных рядов, обрывающиеся в тот момент, когда требуется сделать прогноз на будущее. Для решения задач прогнозирования часто удаётся приспособить методы регрессии или классификации, причём во втором случае речь идёт скорее о задачах принятия решений.

*Обучение без учителя (unsupervised learning).* В этом случае ответы не задаются, и требуется искать зависимости между объектами.

* *Задача кластеризации (clustering)* заключается в том, чтобы сгруппировать объекты в кластеры, используя данные о попарном сходстве объектов. Функционалы качества могут определяться по-разному, например, как отношение средних межкластерных и внутрикластерных расстояний.
* *Задача поиска ассоциативных правил (association rules learning).* Исходные данные представляются в виде признаковых описаний. Требуется найти такие наборы признаков, и такие значения этих признаков, которые особенно часто (неслучайно часто) встречаются в признаковых описаниях объектов.
* *Задача фильтрации выбросов (outliers detection)* — обнаружение в обучающей выборке небольшого числа нетипичных объектов. В некоторых приложениях их поиск является самоцелью (например, обнаружение мошенничества). В других приложениях эти объекты являются следствием ошибок в данных или неточности модели, то есть шумом, мешающим настраивать модель, и должны быть удалены из выборки, см. также робастные методы и одноклассовая классификация.
* *Задача построения доверительной области (quantile estimation)* — области минимального объёма с достаточно гладкой границей, содержащей заданную долю выборки.
* *Задача сокращения размерности (dimensionality reduction)* заключается в том, чтобы по исходным признакам с помощью некоторых функций преобразования перейти к наименьшему числу новых признаков, не потеряв при этом никакой существенной информации об объектах выборки. В классе линейных преобразований наиболее известным примером является метод главных компонент.
* *Задача заполнения пропущенных значений (missing values)* — замена недостающих значений в матрице объекты–признаки их прогнозными значениями.

*Частичное обучение (semi-supervised learning)* занимает промежуточное положение между обучением с учителем и без учителя. Каждый прецедент представляет собой пару «объект, ответ», но ответы известны только на части прецедентов. Пример прикладной задачи — автоматическая рубрикация большого количества текстов при условии, что некоторые из них уже отнесены к каким-то рубрикам.

*Трансдуктивное обучение (transductive learning).* Дана конечная обучающая выборка прецедентов. Требуется по этим частным данным сделать предсказания отностительно других частных данных — тестовой выборки. В отличие от стандартной постановки, здесь не требуется выявлять общую закономерность, поскольку известно, что новых тестовых прецедентов не будет. С другой стороны, появляется возможность улучшить качество предсказаний за счёт анализа всей тестовой выборки целиком, например, путём её кластеризации. Во многих приложениях трансдуктивное обучение практически не отличается от частичного обучения.

*Обучение с подкреплением (reinforcement learning).* Роль объектов играют пары «ситуация, принятое решение», ответами являются значения функционала качества, характеризующего правильность принятых решений (реакцию среды). Как и в задачах прогнозирования, здесь существенную роль играет фактор времени. Примеры прикладных задач: формирование инвестиционных стратегий, автоматическое управление технологическими процессами, самообучение роботов, и т.д.

*Динамическое обучение (online learning)* может быть, как обучением с учителем, так и без учителя. Специфика в том, что прецеденты поступают потоком. Требуется немедленно принимать решение по каждому прецеденту и одновременно доучивать модель зависимости с учётом новых прецедентов. Как и в задачах прогнозирования, здесь существенную роль играет фактор времени.

*Активное обучение (active learning)* отличается тем, что обучаемый имеет возможность самостоятельно назначать следующий прецедент, который станет известен. См. также Планирование экспериментов.

*Метаобучение (meta-learning или learning-to-learn)* отличается тем, что прецедентами являются ранее решённые задачи обучения. Требуется определить, какие из используемых в них эвристик работают более эффективно. Конечная цель — обеспечить постоянное автоматическое совершенствование алгоритма обучения с течением времени.

* *Многозадачное обучение (multi-task learning).* Набор взаимосвязанных или схожих задач обучения решается одновременно, с помощью различных алгоритмов обучения, имеющих схожее внутренне представление. Информация о сходстве задач между собой позволяет более эффективно совершенствовать алгоритм обучения и повышать качество решения основной задачи.
* *Индуктивный перенос (inductive transfer).* Опыт решения отдельных частных задач обучения по прецедентам переносится на решение последующих частных задач обучения. Для формализации и сохранения этого опыта применяются реляционные или иерархические структуры представления знаний.

Выводы

Для решения поставленных задач, наиболее подходящей методикой является использование методов машинного обучения. Так как поставленная задача и входные данные прекрасно вписываются в общую постановку задачи для алгоритмов машинного обучения и представляет из себя стандартную задачу *классификации с обучением с учителем.*

Так, входные данные можно рассматривать как неделимые события прохождения объектов производства по конвейеру с промежуточными результатами о которых известны измерения в контрольных точках линии производства, то есть конечное множество прецедентов (объектов, ситуаций), по каждому из которых собраны (измерены) некоторые данные. И методы машинного обучения, используя эти обработанные данные, позволяют предсказать результат прохождения объектом линии или какой-то её части, что более актуально, то есть построить модель для выявления общих закономерностей, присущих не только этой конкретной выборке, но вообще всем прецедентам, в том числе тем, которые ещё не наблюдались, которые позволяют определить будет ли сбой на производственной линии.

СРЕДСТВА РАЗРАБОТКИ

Для анализа исходных данных, а также разработки последующих моделей будут использованы следующие инструменты для работы с алгоритмами машинного обучения:

* Apache Spark
* Язык программирования python

*Apache Spark*— программный каркас с открытым исходным кодом для реализации распределённой обработки неструктурированных и слабоструктурированных данных, входящий в экосистему проектов Hadoop. В отличие от классического обработчика из ядра Hadoop, реализующего двухуровневую концепцию MapReduce с дисковым хранилищем, использует специализированные примитивы для рекуррентной обработки в оперативной памяти, благодаря чему позволяет получать значительный выигрыш в скорости работы для некоторых классов задач, в частности, возможность многократного доступа к загруженным в память пользовательским данным делает библиотеку привлекательной для алгоритмов машинного обучения.

Проект предоставляет программные интерфейсы для языков Java, Scala, Python, R. Состоит из ядра и нескольких расширений, таких как Spark SQL (позволяет выполнять SQL-запросы над данными), Spark Streaming (надстройка для обработки потоковых данных), Spark MLlib (набор библиотек машинного обучения), GraphX (предназначено для распределённой обработки графов). Может работать как в среде кластера Hadoop под управлением YARN, так и без компонентов ядра Hadoop, поддерживает несколько распределённых систем хранения — HDFS, OpenStack Swift, NoSQL-СУБД Cassandra, Amazon S3.

Так как исходный объем данных достаточно велик (больше 9 Гб), и при дальнейшем использовании системы этот объем будет только увеличиваться, в качестве инструмента разработки был выбран именно Apache Spark. Построенная с его помощью модель может быть перенесена на кластер без изменений.

АНАЛИЗ ДАННЫХ, ГЕНЕРИРУЕМЫХ ДАТЧИКАМИ ПРОИЗВОДСТВЕННОЙ ЛИНИИ

Исходные данные представляют из себя события прохождения объектов производства по конвейеру или его части с промежуточными результатами, о которых известны измерения в контрольных точках линии производства и представлены они в виде таблицы, строи которой это информация движения изделия по линии или её части, а столбцы – измерения в конкретной точке производственной линии. Так как строение производственной линии, а также используемые системы и датчики являются конфиденциальной информацией то, данные анонимизированы, то есть невозможно узнать логическое значение измерений датчика и его название, однако для алгоритма эта информация не является важной, также при таком решение задачи (без привязки к конкретному оборудованию) возможен её перенос на другое производство с небольшими модификациями или без них.

Анализ

Исходные признаки разбиты на три группы: числовые, временные и категориальные. Для итоговой модели нужны все 3 группы, однако, изначально можно работать с ними и проводить анализ по отдельности.

* Числовые признаки – это некоторые измерения различных датчиков, например, датчиков скорости или положения.
* Временные признаки – это время прохождения детали участков производственной линии.
* Категориальные признаки – некоторые атрибуты прохождения детали участков производственной линии. Обычно это заранее заданный диапазон дискретных значений ({П1, П2, П3}).

Для построения хорошей модели необходимо провести анализ данных, выявить зависимости, корреляции, возможно добавить какие-то новые признаки для улучшения качества, либо удалить признаки, вносящие только шум.

Для начала необходимо посмотреть на общие статистики обрабатываемых данных, такие как количество объектов, количество признаков, проценты наполненности данных.

* Количество объектов:
* Количество признаков:

Также необходимо посмотреть, что из себя представляют данные «внутри», то есть увидеть описание некоторой подвыборки данных (объектов). Возьмем выборку в 10% от основного объема объектов:

…

Следующий шаг, это узнать количество пропусков в данных, можно построить следующие диаграммы:

…

Так как поставленная задача – это задача бинарной классификации, то необходимо увидеть распределение объектов по классам. Данное распределение представлено на рис. 1.

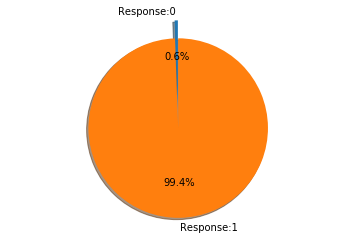


Рис 1. Распределение классов целевой переменной

Из представленного графика видно, что данные сильно несбалансированные, то есть в них присутствует преобладание одного класса объектов со значением целевой переменной равным один.

См Приложение А

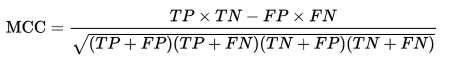
В доработке …

Выбор метрики качества модели

См Приложение А

Выбор метрики качества модели – очень важный аспект построения системы, основанной на методах машинного обучения.

Так как классы сильно несбалансированные, то в качестве метрики качества был выбран «Matthews correlation coefficient», который отлично подходит для данных условий. Формула расчета MCC следующая:



Dsf

ПОСТРОЕНИЕ МОДЕЛИ ПРЕДУПРЖДЕНИЯ СБОЕВ НА ПРОИЗВОДСТВЕННОЙ ЛИНИИ

Описание алгоритма

После анализа данных необходимо выбрать конкретный алгоритм машинного обучения. Для задач классификации при обучении с учителем в настоящее время самым распространенным и эффективным методом является градиентный бустинг над деревьями.

Бустинг (англ. boosting — улучшение) — это процедура последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов. Бустинг представляет собой жадный алгоритм построения композиции алгоритмов.

В течение последних 10 лет бустинг остаётся одним из наиболее популярных методов машинного обучения, наряду с нейронными сетями и машинами опорных векторов. Основные причины — простота, универсальность, гибкость (возможность построения различных модификаций), и, главное, высокая обобщающая способность.

Бустинг над решающими деревьями считается одним из наиболее эффективных методов с точки зрения качества классификации. Во многих экспериментах наблюдалось практически неограниченное уменьшение частоты ошибок на независимой тестовой выборке по мере наращивания композиции. Более того, качество на тестовой выборке часто продолжало улучшаться даже после достижения безошибочного распознавания всей обучающей выборки [1]. Это перевернуло существовавшие долгое время представления о том, что для повышения обобщающей способности необходимо ограничивать сложность алгоритмов. На примере бустинга стало понятно, что хорошим качеством могут обладать сколь угодно сложные композиции, если их правильно настраивать.

Впоследствии феномен бустинга получил теоретическое обоснование. Оказалось, что взвешенное голосование не увеличивает эффективную сложность алгоритма, а лишь сглаживает ответы базовых алгоритмов. Количественные оценки обобщающей способности бустинга формулируются в терминах отступа [1]. Эффективность бустинга объясняется тем, что по мере добавления базовых алгоритмов увеличиваются отступы обучающих объектов. Причём бустинг продолжает раздвигать классы даже после достижения безошибочной классификации обучающей выборки.

Для построения модели будем использовать готовую реализацию алгоритма в свободной библиотеке xgboost. Библиотека xgboost представляет модифицированные, распределенные и оптимизированные алгоритмы градиентного бустинга для классификации и регрессии, а также дополнительные возможности и утилиты для настройки алгоритмов.

См Приложение А.

В доработке …

ЗАКЛЮЧЕНИЕ

В доработке …

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

В доработке …

# ПРИЛОЖЕНИЕ А

*# coding: utf-8  
  
# ### Enviroment preparation  
  
# In[1]:***from** pyspark **import** SparkContext, SparkConf  
**from** pyspark.sql **import** SQLContext, Row  
**from** pyspark.sql.types **import** \*  
**import** pyspark.sql.functions **as** F  
**import** numpy **as** np  
**import** pandas **as** pd  
**import** random  
**import** time **as** t  
**import** os  
**import** re  
**from** subprocess **import** Popen  
**import** matplotlib.pyplot **as** plt  
  
get\_ipython().magic(**u'matplotlib inline'**)  
  
*# In[6]:*ControlSparkCluster(**'stop'**)  
  
  
*# In[2]:***def** ControlSparkCluster(action):  
 **if** action == **'start'**:  
 Popen(**'./start\_cluster.sh'**, shell=**True**)  
 t.sleep(10)  
 **print 'cluster started'  
 elif** action == **'stop'**:  
 Popen(**'./stop\_cluster.sh'**, shell=**True**)  
 sc.stop()  
 t.sleep(3)  
 **print 'cluster stopped'  
 else**:  
 **raise** Exception(**'Unsupported command!'**)  
  
  
*# In[3]:*ControlSparkCluster(**'start'**)  
  
*# In[4]:*conf = SparkConf().setMaster(**"spark://cathome:7077"**).setAppName(**"bosh"**).setAll(([**'spark.executor.cores'**, **'4'**]  
 , [**'spark.driver.cores'**, **'4'**]  
 , [**'spark.driver.memory'**, **'1g'**]  
 , [**'spark.executor.memory'**, **'1g'**]  
 , [**'spark.python.worker.memory'**, **'2g'**]  
 , [**'spark.default.parallelism'**, **'4'**]  
 ))  
  
*# In[5]:*sc = SparkContext(conf=conf)  
sqlContext = SQLContext(sc)  
  
*# ### Data preparation  
  
# In[6]:*dir\_name = **'../../../diploma\_data'***# In[7]:*tr\_num\_file\_name = **'train\_numeric.csv'**tr\_date\_file\_name = **'train\_date.csv'**tr\_cat\_file\_name = **'train\_categorical.csv'***# In[8]:*tr\_num\_data\_file = **'{0}/{1}'**.format(dir\_name, tr\_num\_file\_name)  
tr\_date\_data\_file = **'{0}/{1}'**.format(dir\_name, tr\_date\_file\_name)  
tr\_cat\_data\_file = **'{0}/{1}'**.format(dir\_name, tr\_cat\_file\_name)  
  
*# ### Data analisys/Data transformation  
  
# In[10]:*df\_num2 = df\_num  
df\_date2 = df\_date  
df\_cat2 = df\_cat  
  
*# In[10]:*df\_num2.limit(10).toPandas()  
  
*# In[11]:*df\_date2.limit(10).toPandas()  
  
*# In[12]:*df\_cat2.limit(10).toPandas()  
  
*# In[14]:***print 'numeric dataset row count: '**, df\_num2.count()  
**print 'date dataset row count: '**, df\_date2.count()  
**print 'categorical dataset row count: '**, df\_cat2.count()  
  
*# In[16]:*df\_num2.printSchema()  
df\_date2.printSchema()  
df\_cat2.printSchema()  
  
*# In[22]:*df\_num2.describe(df\_num2.columns[1:5]).show()  
  
*# In[24]:*df\_date2.describe(df\_date2.columns[1:5]).show()  
  
*# In[25]:*df\_cat2.describe(df\_cat2.columns[1:5]).show()  
  
*# In[28]:*cols\_num = [(c.split(**'\_'**)) **for** c **in** df\_num2.columns]  
cols\_num\_df = pd.DataFrame(cols\_num[1:-1], columns=[**'line'**, **'station'**, **'feature'**])  
cols\_num\_df  
cols\_num\_df\_g = cols\_num\_df.groupby([**'line'**, **'station'**], sort=**False**).count()  
cols\_num\_df\_g  
  
*# In[29]:*cols\_date = [(c.split(**'\_'**)) **for** c **in** df\_date2.columns]  
cols\_date\_df = pd.DataFrame(cols\_date[1:], columns=[**'line'**, **'station'**, **'feature'**])  
cols\_date\_df  
cols\_date\_df\_g = cols\_date\_df.groupby([**'line'**, **'station'**], sort=**False**).count()  
cols\_date\_df\_g  
  
*# In[30]:*cols\_cat = [(c.split(**'\_'**)) **for** c **in** df\_cat2.columns]  
cols\_cat\_df = pd.DataFrame(cols\_cat[1:], columns=[**'line'**, **'station'**, **'feature'**])  
cols\_cat\_df  
cols\_cat\_df\_g = cols\_cat\_df.groupby([**'line'**, **'station'**], sort=**False**).count()  
cols\_cat\_df\_g  
  
*# In[75]:*diff\_c = cols\_num\_df.join(cols\_date\_df, on=[**'line'**, **'station'**, **'feature'**], lsuffix=**'\_num'**, rsuffix=**'\_date'**, how=**'outer'**)  
diff\_c  
  
*# In[26]:***def** plt\_hist(counts\_by\_col\_d):  
 hist\_v = map(float, [v **for** k, v **in** sorted(counts\_by\_col\_d.items())])  
 elems = list(xrange(0, len(hist\_v)))  
 plt.figure(figsize=(25, 6))  
 plt.ylabel(**'proc'**)  
 plt.xlabel(**'col id'**)  
 plt.grid(**True**)  
 plt.xlim([-10, len(elems) + 10])  
 width = 1 / 1  
 *# plt.xticks(elems, row\_fileds)* plt.bar(elems, hist\_v, width, color=**"blue"**)  
  
  
*# In[27]:***def** plt\_pipe(counts\_1\_all\_l):  
 labels = **'Response:0'**, **'Response:1'** sizes = (counts\_1\_all\_l[**'count\_1'**], counts\_1\_all\_l[**'count\_all'**])  
 explode = (0.2, 0)  
  
 fig1, ax1 = plt.subplots()  
 ax1.pie(sizes, explode=explode, labels=labels, autopct=**'%1.1f%%'**,  
 shadow=**True**, startangle=90)  
 ax1.axis(**'equal'**)  
  
  
*# In[28]:*counts\_1\_all\_l = df\_num2.select(F.sum(F.when(df\_num2[**'Response'**] == 1, 1).otherwise(0)).alias(**'count\_1'**),  
 F.count(**'\*'**).alias(**'count\_all'**)).collect()  
  
*# In[29]:*plt\_pipe(counts\_1\_all\_l[0].asDict())  
  
*# In[14]:*counts\_by\_col = df\_num2.select(\*[(F.count(c) / F.count(**"\*"**)).alias(c) **for** c **in** df\_num2.columns])  
  
*# In[ ]:*counts\_by\_col.limit(1).toPandas()  
  
*# In[15]:*counts\_by\_col\_l = counts\_by\_col.collect()  
  
*# In[106]:*plt\_hist(counts\_by\_col\_l[0].asDict())  
  
*# In[ ]:*col\_with\_resp = df\_num2.filter(df\_num2.Response == 1).select(  
 \*[(F.count(c) / F.count(**"\*"**)).alias(c) **for** c **in** df\_num2.columns])  
  
*# In[ ]:*col\_with\_resp.limit(100).toPandas()  
  
*# In[ ]:*col\_with\_resp\_l = col\_with\_resp.collect()  
  
*# In[ ]:*plt\_hist(col\_with\_resp\_l[0].asDict())  
  
*# In[16]:*counts\_by\_col3 = df\_date2.select(\*[(F.count(c) / F.count(**"\*"**)).alias(c) **for** c **in** df\_date2.columns])  
  
*# In[ ]:*counts\_by\_col3.limit(1).toPandas()  
  
*# In[17]:*counts\_by\_col\_l3 = counts\_by\_col3.collect()  
  
*# In[117]:*plt\_hist(counts\_by\_col\_l3[0].asDict())  
  
*# In[18]:*counts\_by\_col4 = df\_cat2.select(\*[(F.count(c) / F.count(**"\*"**)).alias(c) **for** c **in** df\_cat2.columns])  
  
*# In[19]:*counts\_by\_col\_l4 = counts\_by\_col4.collect()  
  
*# In[120]:*plt\_hist(counts\_by\_col\_l4[0].asDict())  
  
*# In[20]:*filtered\_num = {k: v **for** k, v **in** counts\_by\_col\_l[0].asDict().iteritems() **if** v > 0.15}  
  
*# In[110]:*plt\_hist(filtered\_num)  
  
*# In[21]:*filtered\_cat = {k: v **for** k, v **in** counts\_by\_col\_l4[0].asDict().iteritems() **if** v > 0.15}  
*# filtered\_cat  
  
  
# In[122]:*plt\_hist(filtered\_cat)  
  
*# In[22]:*filtered\_date = {k: v **for** k, v **in** counts\_by\_col\_l3[0].asDict().iteritems() **if** v > 0.15}  
  
*# In[124]:*plt\_hist(filtered\_date)  
  
*# In[23]:*flt\_df\_num = df\_num2.select(\*filtered\_num.keys())  
flt\_df\_date = df\_date2.select(\*filtered\_date.keys())  
flt\_df\_cat = df\_cat2.select(\*filtered\_cat.keys())  
  
*# In[24]:*flt\_df\_num = flt\_df\_num.drop\_duplicates()  
flt\_df\_date = flt\_df\_date.drop\_duplicates()  
flt\_df\_cat = flt\_df\_cat.drop\_duplicates()  
  
*# In[25]:*replace\_cat\_vals = F.udf(**lambda** val: 1 **if** val == **'T1' else** 9999999, IntegerType())  
flt\_df\_cat = flt\_df\_cat.select(flt\_df\_cat.Id, \*[replace\_cat\_vals(c).alias(c) **for** c **in** flt\_df\_cat.columns **if** c != **'Id'**])  
  
*# In[26]:*flt\_df\_num = flt\_df\_num.na.fill(9999999)  
  
*# In[27]:*flt\_df\_date = flt\_df\_date.na.fill(9999999)  
  
*# In[28]:*full\_df = flt\_df\_num.join(flt\_df\_date, flt\_df\_num.Id == flt\_df\_date.Id, **'inner'**).drop(flt\_df\_date.Id)  
full\_df = full\_df.join(flt\_df\_cat, full\_df.Id == flt\_df\_cat.Id, **'inner'**).drop(flt\_df\_cat.Id)  
  
*# In[29]:*full\_df.write.csv(**'{0}/{1}'**.format(dir\_name, **'full\_df.csv'**), sep=**','**, header=**True**, quote=**''**)  
  
*# In[31]:*full\_df2 = sqlContext.read.csv(**'{0}/{1}'**.format(dir\_name, **'full\_df.csv'**), sep=**','**, header=**True**, inferSchema=**True**)  
  
*# ### Creation of Models  
  
# In[32]:*full\_df3 = full\_df2  
  
*# In[87]:***from** pyspark.ml.classification **import** GBTClassifier, RandomForestClassifier  
**from** pyspark.ml.pipeline **import** Pipeline  
**from** pyspark.ml.linalg **import** Vectors  
**from** pyspark.ml.feature **import** VectorAssembler, PCA  
  
*# In[35]:*features = [c **for** c **in** full\_df3.columns **if** c != **'Id' and** c != **'Response'**]  
  
*# In[36]:*vecAssembler = VectorAssembler(inputCols=features, outputCol=**"features"**)  
  
*# In[38]:*assd = ass.select([**'Id'**, **'Response'**, **'features'**]).drop()  
  
*# In[39]:*assd.limit(5).toPandas()  
*# assd.printSchema()  
  
  
# In[89]:*selector = PCA(inputCol=**"features"**, outputCol=**"pca\_features"**, k=150)  
selectorModel = selector.fit(assd)  
  
*# In[95]:*assd\_pca = selectorModel.transform(assd)  
  
*# In[96]:*assd\_pca = assd\_pca.select(assd\_pca.Id, assd\_pca.Response, assd\_pca.pca\_features.alias(**'features'**))  
  
*# In[97]:*assd\_pca.limit(10).toPandas()  
  
*# In[114]:*assd0 = assd\_pca.filter(assd.Response == 0).sample(**False**, 0.006, 43)  
  
*# In[115]:*assd1 = assd\_pca.filter(assd.Response == 1)  
  
*# In[116]:*assd\_n = assd0.union(assd1)  
  
*# In[117]:*assd\_pca.limit(10).toPandas()  
  
*# In[118]:*tr\_dt, tst\_dt = assd\_pca.randomSplit([0.7, 0.3])  
  
*# In[119]:*rfc = RandomForestClassifier(numTrees=50, maxDepth=5, labelCol=**"Response"**, seed=201)  
  
*# In[125]:*tr\_dt.show()  
  
*# In[60]:*l = [(k, v) **for** k, v **in** enumerate(fti)]  
ls = sorted(l, key=**lambda** r: r[1], reverse=**True**)  
ls  
  
*# In[121]:*pr = model.transform(tst\_dt)  
  
*# In[122]:*pr.show()  
  
*# In[123]:*label\_pred\_list = pr.select(pr[**'Response'**], pr[**'prediction'**]).collect()  
  
  
*# ### Testing Models  
  
# ### Results evaluation  
  
# In[109]:***def** matthews\_corr\_coef(label\_pred\_list):  
 **import** math  
 tp = np.sum(1 **for** r **in** label\_pred\_list **if** r[0] == 1 **and** r[1] == 1)  
 tn = np.sum(1 **for** r **in** label\_pred\_list **if** r[0] == 0 **and** r[1] == 0)  
 fp = np.sum(1 **for** r **in** label\_pred\_list **if** r[0] == 0 **and** r[1] == 1)  
 fn = np.sum(1 **for** r **in** label\_pred\_list **if** r[0] == 1 **and** r[1] == 0)  
 **print "True Positives:"**, tp  
 **print "True Negatives:"**, tn  
 **print "False Positives:"**, fp  
 **print "False Negatives:"**, fn  
 *# print "Total", df.count()* **try**:  
 MCC = ((tp \* tn) - (fp \* fn)) / math.sqrt((tp + fp) \* (tp + fn) \* (tn + fp) \* (tn + fn))  
 **except** ZeroDivisionError:  
 MCC = -1  
 **print "Matthews correlation coefficient: "**, MCC  
  
  
*# In[124]:*matthews\_corr\_coef(label\_pred\_list)