# **Projeto Final de Big Data**

# CEDS-841 *Big Data*

Prof. Dr. Adilson Marques da Cunha

Integrantes

Anderson Baraldo Junior

Leonardo Bernardi de Avila

Luiz Felipe Santana dos Santos

Matheus dos Anjos Inoue

Rômulo Madureira Rodrigues

## Introdução

O primeiro passo para a construção deste trabalho refere-se à precisa definição do que é um “churner”, ou melhor, o que faz de um cliente um ex-cliente. Para o setor bancário, esses são clientes que, deliberadamente, escolhem não renovar seus contratos de empréstimo, resgatam seu empréstimo prematuramente, fecham sua conta bancária na determinada instituição e/ou simplesmente estão dispostos a mudar para outros prestadores de serviços financeiros[1].

A consultoria alemã Mayato diz que as taxas de rotatividade típicas estão entre 10 e 30% para o setor bancário, acarretando não só na perda de receita dos clientes, mas também de outras contribuições indiretas de valor, por exemplo, recomendações em potencial para possíveis outros consumidores.

No Brasil, o potencial de perda de clientes por “churn” aumentou depois da nova resolução do Banco Central sobre a Portabilidade bancaria: CMN - RESOLUÇÃO Nº 4.639, DE 22 DE FEVEREIRO DE 2018, além das pressões de mercado trazidas por novos *players*, como *fintechs* e bancos estrangeiros.

No entanto, a situação encontrada e também trazida pela consultoria é de que a probabilidade de recuperar com sucesso um cliente perdido é alta e com custos comparativamente baixos quando comparado aos custos de conquistar um novo cliente, o custo de recuperar um antigo cliente é em média 1/3 do custo de aquisição de um novo[1].

Não raramente, a questão de se um cliente irá deixar a empresa é frequentemente tratada como um problema de classificação largamente tratado na mineração de dados. Assim, essa problemática pode ser modelada com técnicas de previsão “convencionais”, como árvores de decisão, random forest, procedimento de regressão logística ou redes neurais[1].

Essas técnicas aprendem a diferenciar entre clientes que eventualmente deixaram a empresa (classe 1) e clientes que ainda usufruem dos serviços (classe 0) com base em características gerais de um número de clientes exemplo. O modelo pode posteriormente ser aplicado a clientes não classificados, onde a probabilidade de “churn” – de deixar a empresa - é calculada individualmente com base no padrão de diferenciação aprendido.

O protótipo desenvolvido pelo grupo busca, através de um dataset público encontrado na página do Kaggle com as principais características de clientes atuais e ex-clientes, utilizar ferramentas de Big Data para ingerir, armazenar, processar e acessar os dados de modo a prever a saída de clientes e gerar valor para a instituição em questão.

O principal objetivo é identificar clientes com maior chance de mudar para outra instituição para que as áreas de negócio responsáveis possam então atuar nos diferentes grupos de clientes com ações específicas para estreitar as relações, reduzir o “churn” e garantir o alto grau de satisfação em um “long-term-loyalty”.

Para realizar o protótipo, foi utilizado apenas a ferramenta Spark utilizando uma máquina na nuvem provida pela Databricks e uma interface de programação Python (pySpark), em um ambiente de notebook.

Os dados foram ingeridos a partir de um arquivo .csv baixado da página do Kaggle e então processados e armazenados de maneira distribuída para construção dos modelos de Machine Learning – regressão logística, random forest e gradiente boosting – pelo pacote Spark ML.

Os modelos gerados via Spark alimentaram um recurso de dashboard do ambiente de notebook para simular a operação implementada da solução, com a qual os funcionários do banco responsáveis pelo contato com o cliente final possam então atuar com ações específicas para estreitar as relações, ofertando produtos personalizados e “on-demand”.

Nas seções subsequentes, descreve-se com detalhes o desenvolvimento, resultados e conclusões obtidos neste projeto. Na seção de desenvolvimento, foi endereçado o problema de Big Data resolvido neste estudo com detalhes, incluindo a arquitetura proposta pelo grupo para as quatro fases do Big Data e as diversas maneiras de se modelar a taxa de saída dos clientes do banco em questão.

Na conclusão, pondera-se sobre as principais observações do projeto. Dentre elas, como a arquitetura proposta se comportou com relação a ingestão, armazenamento, processamento e acesso dos dados. Além de observações sobre os benefícios de se usar Spark na realização de tarefas em memória e como suas bibliotecas propiciaram análises detalhadas através de regressão e classificação construídas para prever o potencial de Churn de um determinado cliente.

**Desenvolvimento**

**Procurou-se estudar o comportamento de retenção de clientes, problema conhecido como *Churn customer analysis,* em uma instituição do sistema financeiro nacional*.* Este é um problema muito evidenciado na atualidade devido a pressões regulatórias e de mercado.**

**Devido ao grande número de clientes e consequentemente de dados transacionais dos mesmos, há uma necessidade de processar dados de forma complexa. Para isto, foram utilizadas plataformas de *Big Data* que podem trabalhar com as bases e modelos necessários.**

**Para modelar o problema, foi utilizada uma base sintética disponível na plataforma *Kaggle*, que reúne as seguintes informações dos clientes:**

**- CustomerId: Número de identificação do cliente**

**- Surname: Sobrenome do cliente**

**- Credit Score: Pontuação tirada de agências de crédito**

**- Geography: País no qual o cliente se encontra**

**- Gender: Gênero do cliente**

**- Age: Idade do cliente**

**- Tenure: há quantos anos o cliente está com a empresa**

**- Balance: Saldo do cliente com o banco**

**- Num of products: Quantos produtos financeiros o cliente tem contratado**

**- Has CrCard: o cliente tem cartão de crédito com o banco**

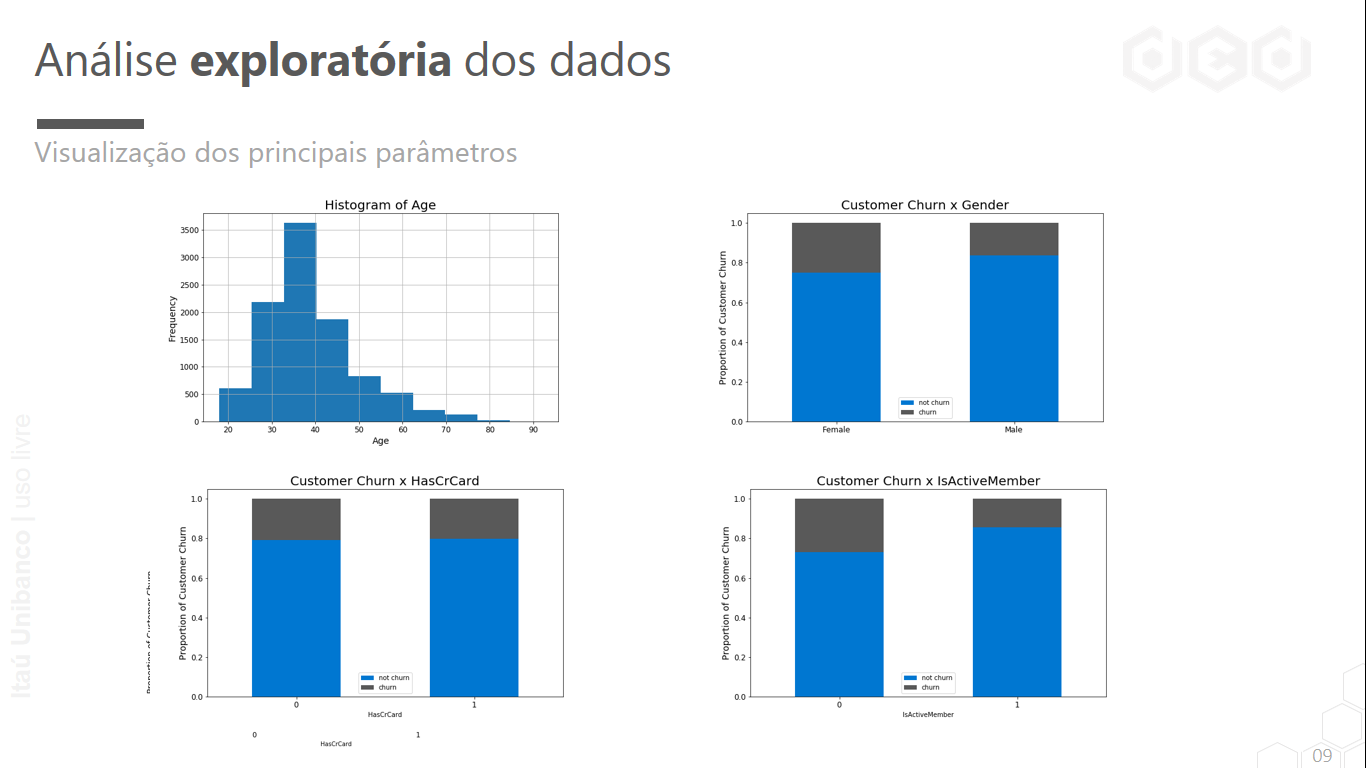
**- Is Active Member: o cliente tem interações frequentes com o banco nos últimos dias**

**- Estimated Salary: Salário anual estimado do cliente**

**E uma variável de saída: Exited, que indica se o cliente deixou a instituição financeira.**

**Análise Descritiva**

**Após a realização da análise dos dados, descobriu-se que a maioria dos clientes da instituição financeira estão incluídos na faixa de 30-50 anos, como pode ser visto na Figura XX. Tal informação poderia ser importante para uma determinada área de negócio que quisesse avaliar a construção e oferecimento de novos produtos e serviços aos seus clientes e, dentre outras abordagens, poderia também servir de critério auxiliar para segmentação.**



**Figura XX**

**Além disso, através da análise gráfica, foi possível obter alguns outros insights da base de clientes:**

* **A proporção de churners é maior dentre clientes do sexo feminino (Customer Churn x Gender);**
* **A posse de produtos cartão de crédito parece não influenciar na questão do cliente se tornar um churner, visto que a proporção de churners dentre pessoas que possuem ou não cartão de crédito é igual (Customer Churn x HasCrCard);**
* **Clientes com maior quantidade de interações com o banco possuem menos chance de deixar a instituição (Customer Churn x IsActiveMember);**

**Sendo assim, esse tipo de informação seria de fundamental importância nas diversas tomadas de decisão intrínsecas ao negócio.**

**Ferramentas de Big Data**

**Existem diversas ferramentas de Big Data disponíveis no mercado, neste tópico serão apresentadas as ferramentas propostas para utilizar na solução do problema.**

**Apache Sqoop**

**Apache Sqoop, capaz de fazer a ingestão de dados massivos provenientes de banco de dados relacionais no sistema de arquivos distribuídos do Hadoop (HDFS). O Sqoop possibilita uma solução de ingestão simples, rápida, de baixo custo de processamento, capaz de realizar a transferência de dados em paralelo, em casos de muitos acessos de escrita no HDFS e pode se conectar com todos os RDBMS mais utilizados no mercado (Oracle, MySQL, Postgre, ..., etc.).**

**Hadoop Distributed File System (HDFS)**

**O HDFS é o sistema padrão de armazenamento de dados distribuídos no ecossistema Hadoop. Ele utiliza a infraestrutura do Hadoop para prover alta performance de acesso a dados em clusters de hardware commodities, permitindo alta escalabilidade de armazenamento necessária em aplicações de Big Data Analytics. Possui grande flexibilidade, podendo armazenar tanto dados estruturados como não estruturados.**

**Apache Spark**

**O Spark é uma ferramenta de processamento alternativa ao Hadoop MapReduce que possui alta performance em paralelismo de tarefas e tempo de execução, chegando a ser de 10 a 100 vezes mais rápido que o tradicional MapReduce e é capaz de realizar o processamento massivo de dados em clusters.**

**Essa superioridade de performance se deve a forma como o Spark organiza suas tarefas, utilizando o paradigma de grafos direcionais acíclicos (DAG), o Spark escalona suas tarefas com o intuito de otimizá-las, inclusive reaproveitando processamento já realizado. Ao contrário do MapReduce que precisa persistir os resultados de suas operações em disco, o Spark pode armazenar resultados em memória, o que permite rápido acesso em etapas posteriores.**

**Existem dois tipos de operações no Spark, transformação e ação. Devido ao paradigma DAG o Spark pode fazer uso da avaliação tardia (Lazy Evaluation) no seu processamento, desta forma, as transformações são escalonadas e distribuídas nos clusters, porém a execução só é efetivamente realizada quando uma ação é solicitada. Desta forma pode-se fazer sequências (pipelines) de operações de transformação de dados na programação. O Spark é programado em Scala, uma linguagem que segue o paradigma funcional.**

**Outra grande vantagem do Spark é a disponibilidade de APIs para linguagens de programação populares em Data Science como Python, R, SQL e Java, o que permite utilizar outras bibliotecas dessas linguagens em projetos garantindo grande flexibilidade na programação.**

**O Spark possui uma variedade de bibliotecas próprias, com destaque às bibliotecas de Machine Learning (SparkML), que permitem o processamento de algoritmos de ML de forma distribuída e eficiente, uma grande vantagem para o deploy de modelos.**

**Plataforma Databricks**

**A Databricks é uma empresa fundada pelos criadores do Apache Spark, oriundos da Universidade da Califórnia, Berkeley. Eles disponibilizam como serviço uma plataforma unificada de Data Science que tem como elemento principal o Spark, para que times de cientistas de dados possam construir soluções de Big Data Analytics. Nesta plataforma, a programação é feita por uma interface via notebooks *Zeppelin*, que torna possível o desenvolvimento e a análise dos dados on the fly.**

**Arquitetura**

**Para realizar um projeto de Data Science escalável, é necessário desenvolver uma arquitetura de Big Data Analytics que atenda com completeza os requisitos de performance do projeto, definindo todas as fases de processamento de Big Data para garantir que o problema será resolvido de forma eficaz. A arquitetura escolhida, portanto, segue o esquema da Figura 1.**



**Para a ingestão de dados é sugerido o uso da ferramenta Apache Sqoop que possui grande eficiência e baixo custo em transferir dados de bancos de dados relacionais (RDBMS) para o HDFS, podendo trabalhar com os vários tipos de RDBMS disponíveis no mercado. Para o armazenamento é utilizado o HDFS, pois permite o armazenamento distribuído e escalável dos *Big Data sets*. Já para processamento foi utilizado o Apache Spark através de sua API de python (Pyspark), onde foram construídos os modelos de aprendizado de máquina utilizando a biblioteca SparkML. Todos esses recursos são providos pela plataforma Databricks, que possibilita também a criação de dashboards interativos, necessários para a fase de acesso aos dados e deploy dos modelos.**

**Modelos de Machine Learning**

**Como a base possui dados rotulados, isso permite o uso de modelos de aprendizado de máquina supervisionados para a classificação: se os clientes irão ou não encerrar o relacionamento com o banco. Portanto, foram criados diferentes modelos de classificação utilizando o SparkML, com o objetivo de obter o melhor resultado na predição de churn. Os algoritmos testados foram:**

**- REGRESSÃO LOGISTICA**

**- RANDOM FOREST**

**- GRADIENT BOOSTING**

**Resultados dos modelos de Machine Learning**

**A tabela \ref{} contém a avaliação dos modelos utilizando as métricas de área abaixo da curva (AUC), acurácia e recall.**

**O modelo que utiliza o algoritmo de Gradient Boosting conseguiu a melhor performance no geral.**

**Modelo de Recomendação**

**Para a implementação do modelo de recomendação, foi utilizada a técnica de *collaborative filtering*, que utiliza as informações da base completa como forma de recomendar itens similares a clientes com padrões similares. Assim, este método foi realizado com o auxílio da técnica SVD (Singular Vector Decomposition), na qual a matriz das preferências conhecidas é decomposta em outras e, quando é reconstruída, as posições nas quais era indicado desconhecimento do cliente em relação ao produto mostram valores diferentes de zero, o que pode ser interpretado como uma pontuação de provável interesse do cliente a este produto.**

**Protótipo de Projeto de Big Data**

**O objetivo do protótipo criado no projeto é, por meio de ferramentas de Big Data, entender como determinados atributos influenciam no cancelamento do serviço financeiro por um cliente e produzir um protótipo de interface no qual um gerente possa atuar para identificar clientes com alto potencial de *churn*, além de prover produtos diferenciados de modo a manter o interesse do cliente pela empresa.**

**Para isto, foram desenvolvidas duas hipótese: a primeira delas é a de que o comportamento de *churn* pode ser previsto ao se utilizar uma base com atributos relevates. A segunda hipótese é a de que, com as ferramentas de *big data* apresentadas anteriormente, é possível desenvolver um protótipo que funcione de modo rápido e gere valor à empresa.**

**Desta forma, ao testar as hipóteses, foi possível atuar em *Big Data Analytics* de forma preditiva, ao criar um modelo que dá uma probabilidade de o cliente deixar a empresa; de forma descritiva, ao criar as visualizações das bases de clientes e verificar diferenças entre substratos distintos quanto à taxa de *churn* e, por fim, houve a atuação de forma prescritiva, ao se criar um sistema de recomendação que pode sugerir produtos correlacionados aos hábitos de consumo dos clientes, o que melhora a relação deste com a empresa e impede sua saída.**

**Através da análise descritiva, pôde-se notar que certos grupos de clientes possuem maior taxa de *Churn*, o que indica a falta de produtos pensados nestes segmentos ou atendimento ineficiente. Além disto,**

**Quanto às contribuições dos integrantes, o grupo produziu uma apresentação coesa sobre os conceitos de *Big Data*, produzindo ricas discussões que resultaram no protótipo apresentado. Assim, a elaboração do projeto foi uma atividade colaborativa.**

**Por fim, o desenvolvimento do trabalho trouxe valores de curto e longo prazo: a curto prazo, o protótipo trouxe a oportunidade de o grupo aprender sobre os conceitos e técnicas de *Big Data*, podendo aplicá-las em favor do banco. A médio e longo prazo, os conhecimentos adquiridos sobre *churn* e *recommender systems* podem ser aplicados em produtos futuros e discussões de negócio do banco.**

1. **O que você descobriu, após realizar alguns tipos de Análises de Dados?**

**III – CONCLUSÃO**

Para concluir, através da experimentação realizada como Projeto Final do curso de Fundamentos de Big Data foi possível verificar que a arquitetura proposta para a ingestão, armazenamento, processamento e acesso dos dados se mostrou eficaz para um Dataset pequeno e está pronta para ser escalada conforme necessidade e utilização de bases de dados reais, que possuam as características de Big Data.

Além disso, o Spark se mostrou eficiente e flexível na realização de tarefas em memória e suas bibliotecas propiciaram análises detalhadas em Machine Learning do problema investigado, onde, através de ferramentas “convencionais” de regressão e classificação foi possível prever o potencial de Churn dos clientes com acurácia satisfatória.

Com essa informação em mãos, foi possível realizar a implementação de um dashboard que pode ser utilizado por funcionários do banco para adquirir conhecimento sobre os clientes e sugerir produtos que sejam satisfatórios e estreitem as relações instituição-cliente e gerem valor através de um sistema de recomendação efetivo.

Dentro das recomendações propostas para este trabalho, algumas são importantes devido a estarem presentes na fase de pré-processamento dos dados, sendo então capazes de influenciar no desempenho dos modelos de machine learning utilizados durante a aplicação. A primeira recomendação seria a realização do balanceamento dos dados, através da aplicação da metodologia SMOTE. O balanceamento seria de extrema importância, porque problemas de classificação que utilizam dados desbalanceados tendem para a classe que majoritária e, com a aplicação desse método, os resultados alcançados seriam mais satisfatórios.

A segunda recomendação envolve a utilização de ferramentas específicas de visualização de dados. Para o dataset considerado, tem-se que o Tableau seria uma boa opção, pois permite a rápida criação de visualizações interativas, é capaz de processar alguns milhões de linhas com certa facilidade permitindo a construção de conhecimento a partir da exploração da base de dados.

Para finalizar, um terceiro ponto de recomendação é o de se realizar o aprimoramento do sistema de recomendação personalizado para os clientes com alto potencial de churn, de forma a garantir que cada sugestão do sistema trate o cliente como se fosse único, o que melhora o nível de satisfação com relação aos serviços oferecidos.

Dado o alto potencial de resultados que esse tipo de análise pode trazer, a principal sugestão de trabalhos futuros é a implementação real de um sistema de predição de churn dentro do banco com a utilização de dados reais com características de Big Data.

Outro ponto importante é a necessidade da integração dos diversos bancos de dados disponíveis, de forma a permitir a construção de modelos completos de predição de churn para tomadas de decisão. Além disso, esse sistema também permitiria diagnósticos que informassem às áreas de negócio quais os principais fatores responsáveis por levar os diversos grupos de clientes a deixar o banco.

Por fim, a utilização desses dados em modelos de recomendação pode ser uma ferramenta auxiliar para prevenir o acontecimento do churn, visto que o banco será capaz de tomar ações pró-ativas para atender às necessidades de cada cliente.

**REFERÊNCIAS**

Referências Bibliográficas:

[1] Churn Analysis for a Direct Bank – with an Eye for the Right Timing - Mayato

[2] Churn Bank Dataset - Fonte: <https://www.kaggle.com/hj5992/bank-churn-modelling>. Acessado em 28/09/2018

[3] Gabriel Moreira - Recommender Systems in Python 101 - <https://www.kaggle.com/gspmoreira/recommender-systems-in-python-101>. Acessado em 28/09/2018

[4] <https://data-flair.training/blogs/dag-in-apache-spark/>. Acessado em 28/09/2018

[5] Notas de aula Introdução ao Big Data do Prof. Adilson Cunha

[6] Classification and Regression on Spark - <https://spark.apache.org/docs/2.1.0/ml-classification-regression.html>. Acessado em 28/09/2018

[7] Evaluation metrics on Spark - <https://spark.apache.org/docs/2.1.0/mllib-evaluation-metrics.html>.  Acessado em 28/09/2018