Universidade Federal de Pernambuco
Centro de Informática

Graduação em Ciência da Computação

O processo CRISP-DM aplicado na construção de uma solução para Análise de Risco de Crédito

Proposta de Trabalho de Graduação

Aluno: Milton Vasconcelos da Gama Neto (mvgn@cin.ufpe.br)
Orientador: Germano Crispim Vasconcelos (gcv@cin.ufpe.br)
Área: Ciência dos Dados

Recife, Agosto de 2018

Resumo

As empresas coletam uma quantidade imensa de dados dos seus usuários, e procuram utilizar essa massa de dados com objetivo de aumentar os lucros, criar insights, melhorar os serviços, entre outros fins. Devido a grande quantidade de dados, se torna inviável a análise e reconhecimento de padrões ao olho nu. A Ciência dos Dados é responsável pela sistematização da análise a partir de técnicas estatísticas, computacionais, aprendizagem de máquina e do conhecimento no domínio do problema. As informações e conhecimentos extraídos dos dados são resultado de um longo processo de Descoberta de Conhecimento em Base de Dados, conhecido como KDD (do inglês Knowledge Discovery in Database). Considerado uma implementação desse processo, o CRISP-DM é a metodologia mais utilizada no mercado, incluindo fases de entendimento do negócio e implantação da solução. Neste trabalho, será detalhado cada etapa desse processo e aplicá-lo na construção de uma solução de Ciência dos Dados para um problema de análise de risco de crédito. No final, será realizada a implantação de um sistema para classificação, utilizando aprendizagem de máquina, e apresentado os resultados do processo, como a avaliação do modelo.

Introdução

Vivendo a era dos dados (HAN; KAMBER; PEI, 2011), onde milhões de informações são geradas por segundo através smartphones, redes sociais, internet, sensores, entre outras fontes. As empresas, startups e pesquisadores, estão buscando utilizar esses dados da melhor maneira. Com objetivos diversos, principalmente em aumentar os lucros, melhorar os processos ou gerar insights para novas soluções.

Os dados são considerados como ouro do século XXI, mas não basta ter-los, é necessário saber o que fazer com eles, como realmente transformar-los em algo valioso. Nesse contexto que entra a Ciência dos Dados, um termo antigo que passou a ser amplamente utilizado com o surgimento de Big Data e o desenvolvimento de Machine Learning.

Ciência dos Dados é a sistematização da análise de dados, que devido ao grande volume desses, se torna inviável a realização dessa tarefa por um humano. Então, são utilizadas técnicas estatísticas, matemáticas e computacionais para obter informação e conhecimento a partir dos dados através de um conjunto de princípios fundamentais (Provost; Fawcett). Segundo Hayashi e Chikio, em *Data Science, Classification, and Related Methods,* Ciência dos Dados é o "conceito para unificar estatística, análise de dados, aprendizagem de máquina e seus métodos relacionados". Por ser uma área interdisciplinar, o perfil de um cientista de dados envolve conhecimento e habilidades de diferentes áreas, Drew Conway mostra através de um Diagrama de Venn (Figura 1) a interdisciplinaridade e interseção que representa essa área. Onde é destacada a importância da especialidade no domínio (Substantive Expertise), para definir os alvos corretos, levantar questionamentos e hipóteses sobre o problema específico.

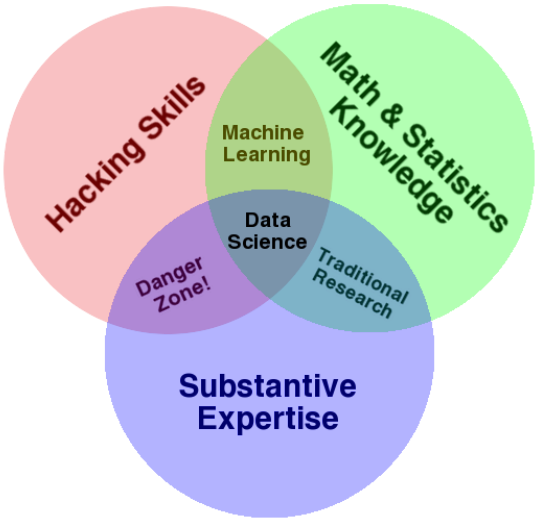

Figura 1: Diagrama de Venn de Drew Conway sobre Ciência dos Dados

Os padrões identificados nos dados, interpretações ou avaliações são o resultado final de um longo processo, que é dividido em cinco fases na Descoberta de Conhecimento em base de dados (KDD, do inglês *Knowledge Discovery in Database*) (Fayyad, 1996). Esse processo sistemático define que o conhecimento é obtido após a realização da seleção dos dados, pré-processamento, transformação, mineração de dados e avaliação. O entendimento dessa jornada é importante para compreender que a partir da obtenção dos dados existe um longo processo, que consome bastante tempo, até chegar em um resultado final.

Das metodologias utilizadas em projetos de Ciência dos Dados, CRISP-DM (*Cross Industry Standard Process for Data Mining*) (Chapman et al, 2000) é a mais utilizada (Piatetsky, 2014). Esse processo, ilustrado na Figura 2, consiste de 6 etapas (entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação) e pode ser considerado uma implementação do KDD [Ana Azevedo, M.F. Santos]. Por ser o mais utilizado e incluir fases como entendimento do negócio e implantação, é um processo mais completo comparado a outros, como SEMMA (*Sample, Explore, Modify, Model, Assess*) (SAS Institute, 2014). Por isso, essa metodologia foi adotada para o trabalho e suas fases serão detalhadas e aplicadas numa base de dados real.

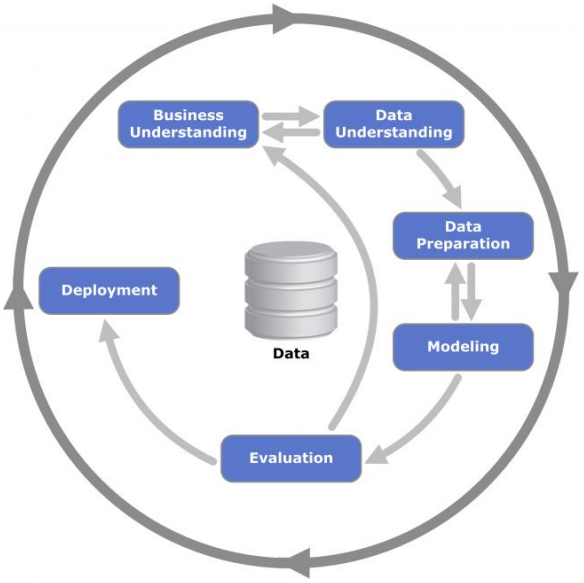

Figura 2: Processo CRISP-DM

A aplicação da metodologia, será feita em uma base de dados para análise de risco de crédito. O processo começará da preparação dos dados, terceira fase do CRISP-DM, pois a empresa que disponibilizou a base já realizou a seleção, definiu o problema e o alvo. Com aproximadamente 400 mil registros de diferentes indivíduos, será construído um classificador, através de aprendizagem de máquina, para definir se ocorrerá inadimplência dado um limiar (risco) na aprovação do crédito.

Objetivos

O objetivo deste trabalho é detalhar as etapas de um projeto de Ciência dos Dados, através do processo CRISP-DM (Cross Industry Standard Process for Data Mining). E aplicar essa metodologia, a partir da fase de pré-processamento, na construção de uma solução para análise de crédito. A solução é modelada em um problema de aprendizagem supervisionada do tipo classificação. Com a realização de experimentos, será selecionado o melhor modelo, baseado em técnicas de avaliação de classificadores, e realizar a implantação deste, simulando o sistema em produção.

Referências

Azevedo, Ana Isabel Rojão Lourenço, and Manuel Filipe Santos. "KDD, SEMMA and CRISP-DM: a parallel overview." IADS-DM (2008).

Chapman, P. et al, 2000. CRISP-DM 1.0 - Step-by-step data mining guide.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17.3 (1996): 37.

Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

Hayashi, Chikio. "What is data science? Fundamental concepts and a heuristic example." Data Science, Classification, and Related Methods. Springer, Tokyo, 1998. 40-51.

Piatetsky, Gregory. "CRISP-DM, still the top methodology for analytics, data mining, or data science projects." KDD News (2014).

Provost, Foster, and Tom Fawcett. Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.", 2013.

SAS Enterprise Miner – SEMMA. SAS Institute, 2014 [online] available: http://www.sas.com/technologies/analytics/datamining/miner/semma.html (September 2014.)

Cronograma

Atividades	Ago		Set			Out			Nov			Dez	
Revisão bibliográfica													
Elaboração da proposta													
Preparação dos dados													
Experimentação													
Avaliação e Implantação													
Escrita do documento final													
Preparação da defesa													

Possíveis Avaliadores

- Ricardo Bastos Prudêncio
- Cleber Zanchettin
- Renato Vimieiro

Assinaturas

_____

Milton Vasconcelos da Gama Neto
(Aluno)

_____

Germano Crispim Vasconcelos
(Orientador)