Como a GDRP Poderá Afetar as Práticas em Ciência de Dados

Uma medida aprovada pelo Parlamento Europeu que define o nível de proteção dos dados de consumidores/clientes de Empresas europeias através de regras de segurança de dados, hoje essas empresas são chamadas controladoras de dados.

Apesar da maior parte dos comentários serem sobre como as regras do GDPR afetam a descoberta e manipulação de Informação pessoalmente identificável (Personally Identifiable information -PII), as regras também afetam as práticas e processos em Ciência de Dados, o que será o foco da discussão do presente artigo.

As regras da GDPR impõem limites nas seguintes práticas em Data Science[1]:

Limitar o processamento de dados pessoais e construção de perfil: Qualquer forma de tratamento automatizado de dados pessoais que consiste na utilização de dados pessoais para avaliar determinados aspectos pessoais relacionados com uma pessoa, em particular, para analisar ou prever aspectos relativos ao desempenho dessa pessoa no trabalho, situação econômica, saúde, preferências pessoais, interesses, confiabilidade, comportamento, localização ou movimentos deverá ser notificada a pessoa e será permitida apenas com a permissão da mesma. E o processamento de dados pessoais será permitida apenas com o propósito específico do negócio original. Ou seja proíbem o uso dos dados para obter qualquer outro tipo de resultado. E apenas com permissão, será permitido usar dados pessoais para qualquer tipo de sistema de predição ou tomadas de decisão automática. O uso de dados anônimos são permitidos, o que fará muitos recorrerem a ferramentas de anonimato dos dados.

Garantir "O direito de explicação" dos sistemas de tomadas de decisões automáticas de IA, isto é utilização de métodos ou técnicas de IA de fácil explicação — Empresas que usam sistemas automatizados devem explicar como estão utilizando os algoritmos para tomada de decisões. Essa "regra" tem dado bastante trabalho aos advogados especialistas, pois parece estar mal definido no conjunto de regras. Enquanto alguns dizem que esta não é regra mas se for trará benefícios outros exageram em falar que métodos de aprendizagem "caixa preta" como Deep Learning se tornarão ilegais, como no artigo intitulado "GDPR tornará aprendizagem de máquina uma prática ilegal"[2]. Especialistas do Instituto Alan Turing dizem que não há motivo pra tanto mau entendimento pois parece estar óbvio que o tal "direito de explicação" trará benefícios pois teremos o direito de questionar sobre como as decisões estão sendo tomadas sobre nós, além da oportunidade de pesquisa[3]. Enquanto as empresas que armazenam e processam nossos dados não querem entregar a fórmula secreta de seus algoritmos pois isso os farão perder posição de mercado.

Evitar viés e discriminação: Algoritmos ficam tendenciosos e começam a influenciar de forma estreita, impondo preferencias raciais, sexuais, políticas, sociais, etc. A GDPR proíbe expressivamente o uso de características pessoais como idade, raça, gênero em decisões automatizadas. Os cientistas de dados também devem adotar medidas afirmativas para confirmar se os dados que usam quando desenvolvem modelos preditivos são precisos; "Garbage In/Garbage Out", ou GIGO, não é uma defesa. Eles também devem considerar se os dados de treinamento tendenciosos sobre os resultados anteriores podem influenciar os modelos. Como resultado, os cientistas de dados precisarão se preocupar com a linhagem de dados para rastrear o fluxo de dados em todas as etapas de processamento, da origem ao destino. O GDPR também gerará maior preocupação com a reprodutibilidade ou a capacidade de replicar com precisão um projeto de modelagem preditiva.

O conjunto de regras da GDPR parecem deixar a vida dos cientistas de dados um tanto "preocupantes".