IBM Global Business Services Brasil

Desafio Advanced Analytics
Análise Exploratória, Modelagem e Otimização

1. Introdução

Parabéns! Você foi escolhido como potencial candidato para trabalhar em projetos de consultoria em Advanced Analytics na IBM. Projetos desta natureza são parecidos com o que vemos nos cases de consultorias estratégicas, onde para resolver um problema formulamos hipóteses e utilizamos dados para confirmá-las ou refutá-las, entregando ao final do trabalho recomendações que geram impacto de negócio para o cliente. Porém, a diferença fundamental é que os dados disponíveis no nosso caso em geral não são facilmente analisados sem a ajuda de ferramentas computacionais. Portanto, temos que usar técnicas de inferência estatística, aprendizado de máquina e inteligência artificial para modelar sistemas complexos como dispersão aérea de contaminantes industriais, falhas catastróficas em máquinas de centenas de toneladas, ocorrência de acidentes de trabalho, fraude em seguradoras, recomendação personalizada de produtos, previsão de demanda por suprimentos, entre outros.

Para trabalhar nestes projetos, o consultor precisa apresentar um conjunto de habilidades específico, onde pensar analiticamente (e fora da caixa) é a habilidade mais importante. Esperamos que este desafio sirva como uma pequena e divertida amostra dos problemas que enfrentamos no cotidiano (ele é baseado em um caso real) e permita a demonstração das suas habilidades como cientista de dados. Vamos lá!

2. Problema

Fomos abordados por um cliente que tem um processo industrial que transforma chocolate em bombons. Uma visão geral do processo é dada na Figura 1. A saída do processo, representada pela variável `PESO_BOMBOM`, pode ser interpretada como o "peso"(massa) médio do lote de bombons produzido. Além disso, temos 3 variáveis de processo: `QTD_CHOC`, `VAR_1` e `VAR_2`. A única variável que podemos controlar é `QTD_CHOC`, a quantidade de chocolate na entrada do processo. As outras duas variáveis não são controláveis por motivos diversos e não têm significado físico evidente.

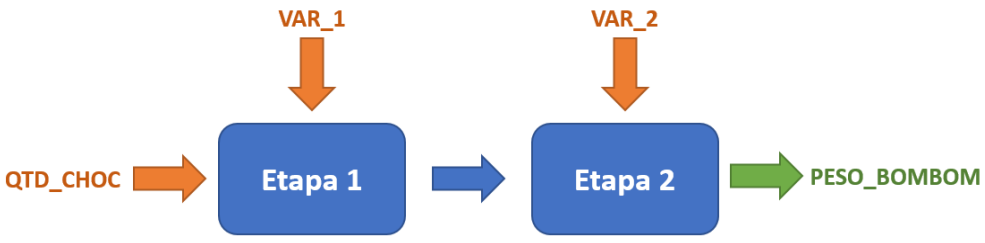

Figura 1. Visão geral do processo de fabricação de bombons.

O cliente mostrou preocupação com o controle de qualidade do seu processo industrial. O peso de um bombom deve ser de 10.0 gramas. Normas de qualidade estabelecem um

limite de variação do peso do bombom de 10.0% somente para baixo. Ou seja, se um bombom pesar menos que 9.0 gramas, ele é considerado não-conforme e descartado. Não há nenhuma norma estabelecendo um limite superior para o peso dos bombons, porém os custos de produção aumentam à medida que bombons acima do peso são produzidos. O gerente da qualidade da empresa compartilhou com a equipe uma métrica de custo empregada para medir a eficiência do processo, mostrada na Figura 2. O custo é mínimo no peso esperado dos bombons de 10.0 gramas e aumenta abruptamente à medida que o peso se aproxima de 9.0 gramas, para modelar o descarte de bombons. Além disso, o custo também aumenta com o excesso de peso, apesar deste aumento ser mais suave.

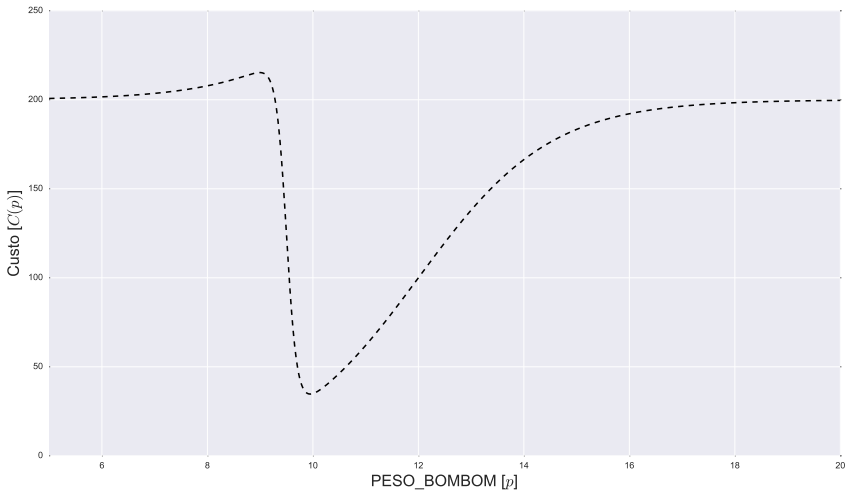

Figura 2. Métrica de custo de processo empregada pelo cliente.

A métrica de custo $C(p)$ se relaciona com o peso p dos bombons segundo a seguinte fórmula:

$$C(p) = \frac{200}{1 + e^{10 \cdot (p - 9.5)}} + \frac{200}{1 + e^{-0.8 \cdot (p - 12)}}$$

O grande problema enfrentado pelo cliente é a variabilidade no processo. Nas palavras do gerente da qualidade:

Estamos tendo dificuldade para modelar nosso processo industrial. As diferentes variáveis de processo impactam de maneira diferente o peso final dos bombons, além de introduzir diferentes níveis de incerteza no processo. Não estamos conseguindo lidar com esta incerteza.

Para ilustrar seu ponto, o gerente mostrou duas comparações. Na Tabela 1 temos o valor da variável `PESO_BOMBOM` para a mesma quantidade da variável `QTD_CHOC` mas para diferentes valores de `VAR_1` e `VAR_2`, ilustrando o impacto das variáveis não controláveis no processo. Além disso, para ilustrar a incerteza no processo, o cliente exibiu o valor da variável `PESO_BOMBOM` para diferentes rodadas de produção com as variáveis `QTD_CHOC`, `VAR_1` e `VAR_2` fixas, mostrado na Tabela 2.

QTD_CHOC	VAR_1	VAR_2	PESO_BOMBOM
300	2.3	A	12.54 g
300	3.0	B	10.09 g
300	1.2	C	8.72 g

Tabela 1. Impacto das variáveis não controláveis no processo.

QTD_CHOC	VAR_1	VAR_2	PESO_BOMBOM
300	2.0	A	12.16 g
300	2.0	A	11.32 g
300	2.0	A	10.80 g

Tabela 2. Variabilidade no processo para mesmo conjunto de variáveis.

O objetivo do cliente com este projeto é identificar pontos de melhora no processo, além de obter um plano de produção (valores ótimos da quantidade de chocolate QTD_CHOC) visando a minimização da métrica de custo. Foi fornecida uma tabela com 500 registros de produção, mapeando o peso do bombom para diversas condições de produção, disponível em anexo.

3. Resolução do problema

A resolução do problema deve ser estruturada conforme as 4 etapas usuais de um projeto de Advanced Analytics: Análise Descritiva, Análise Diagnóstica, Análise Preditiva e Análise Prescritiva. A seguir descrevemos o que é esperado em cada uma das etapas, sugerindo um roteiro de resolução.

3.1. Análise Descritiva

Análise descritiva é a utilização de mineração e visualização de dados para responder a pergunta: **o que aconteceu?** Sugestões para esta entrega:

Explore os dados identificando quais variáveis são contínuas/categóricas e suas distribuições.

Visualize as relações entre as variáveis, particularmente entre as variáveis independentes e a variável alvo.

Apresente uma visão geral do processo, identificando pontos-problema e comportamentos interessantes. Tente **contar uma história** com os dados.

3.2. Análise Diagnóstica

Na **Análise Diagnóstica** identificamos correlações/causalidades e realizamos análise exploratória para responder a pergunta: **por que aconteceu?** Sugestões:

Identifique o problema de negócio e procure nos dados situações que gerem este problema.

Exponha uma conclusão que foi retirada da análise dos dados e tem potencial de geração de valor para o negócio.

Recomende uma possível ação para melhorar o resultado do processo.

3.3. Análise Preditiva

Na **Análise Preditiva** utilizamos inferência estatística e modelos preditivos para responder: **o que vai acontecer?** Sugestões:

Crie um modelo que represente o processo industrial.

Apresente a intuição que suporta o modelo escolhido.

Comprove a assertividade deste modelo através de experimentos.

Preencha os valores em branco da variável `PESO_BOMBOM` no documento `analise-preditiva.xlsx`. As previsões serão avaliadas através da métrica MSE (*Mean Squared Error*).

3.4. Análise Prescritiva

Na **Análise Prescritiva** utilizamos simulação e otimização computacional para responder: **o que fazer?**

Crie um algoritmo que escolha valores ótimos da variável controlável `QTD_CHOC` visando a minimização da métrica de custo do cliente.

Controle a incerteza do processo, estabelecendo um compromisso entre o risco de buscar o custo mínimo e gerar bombons não-conformes, com o aumento do custo associado a gerar bombons acima do peso.

Preencha os valores em branco da variável `QTD_CHOC` no documento `analise-prescritiva.xlsx` com os valores ótimos encontrados. As prescrições serão avaliadas através de simulações repetidas do processo verdadeiro. O custo médio gerado pelos valores da solução será avaliado.

4. Considerações finais

A escolha de ferramenta para conduzir suas análises é livre e não será julgada. Seguem algumas sugestões: Microsoft Excel, SPSS, SAS, Stata e linguagens MATLAB, R, Python, Julia, Ruby e Java. Note que as 4 etapas da solução seguem uma sequência lógica: uma boa análise descritiva facilitará a identificação de pontos problema para obter recomendações na análise diagnóstica; uma boa análise diagnóstica aponta quais transformações e técnicas devem ser usadas para gerar um bom modelo preditivo; uma boa análise diagnóstica e um bom modelo preditivo permitem prescrever ações e executar simulações do sistema para criar hipóteses de otimização. Não é obrigatória a realização de todas as etapas da solução. É preferível realizar as duas primeiras etapas com mais cuidado para obter bons resultados ao invés de realizar as quatro etapas com pouco rigor.

Dicas:

- Você não precisa construir um modelo preditivo para preencher as prescrições na fase da análise prescritiva. Se a análise diagnóstica for bem feita, você já poderá recomendar mudanças na `QTD_CHOC` que gerem redução de custo.
- Dê atenção especial às duas primeiras etapas da resolução. Muitas vezes clientes somente desejam entender um pouco mais seus processos. Use a criatividade e organize suas descobertas como uma pequena história baseada em dados.

- Vemos com bons olhos se o candidato lançar mão de técnicas avançadas de inferência estatística. Porém, não se preocupe em fazer nada muito complicado, se não tiver experiência em modelagem. A avaliação será baseada na sua capacidade analítica.
- Caso não consiga realizar as etapas de modelagem por alguma dificuldade, escreva por extenso como você solucionaria o problema. Um boa ideia de solução, no nosso dia a dia, pode mudar o resultado de um projeto.

Finalmente, bom trabalho! Happy modeling!