Relatório

1. INTRODUÇÃO

Este relatório foi desenvolvido com objetivo de apresentar análises e resultados referentes aos fornecidos no seguinte link. No decorrer do relatório serão apresentados métodos, procedimentos, raciocínios e conclusões tomadas com base na análise dos dados disponibilizados. Para realização das análises foi utilizada basicamente a linguagem de programção Python no ambiente interativo do Jupyter. Todas análises descritas neste relatório podem ser acompanhadas pelos notebooks disponibilizados no Github do autor.

2. ANÁLISE EXPLORATÓRIA

Essa etapa serviu para tomar conhecimento da natureza dos dados a serem modelados e o formato que estes estão distribuídos na amostra. O conjunto de dados é referente à análises de RH. Foi solicitado no desafio que fossem enumerados possíveis problemas passíveis de resolução utilizando machine learning sobre o dataset, ao tomar conhecimento dos dados decidimos selecionar 3 features de interesse que estão contidas no conjunto de dados que se mostraram interessantes para serem utilizadas como target, elas são:

1. **last_evaluation -** A pontuação do funcionário na última avaliação.
2. **left_company** - Informação se o funcionário deixou a companhia
3. **will_relocate** - Informação se o funcionário está disposto a se mudar.

A variável 1 é contínua e tem um intervalo aparente 0 a 1, porém o valor mínimo encontrado no dataset é de 0.36. A distribuição de frequência dessa variável tem uma leve assimetria à esquerda devido a moda ser única e menor que a média e mediana. Decidimos escolher essa variável como possível target pois pode haver interesse em automatizar a avaliação dos funcionários com base em características específicas. Mais a frente explicamos como realizamos a modelagem do problema de regressão desta variável.

A variável 2 é binária e sua proporção está desbalanceada no dataset onde 76% dos dados são da classe 0 enquanto apenas 24% da classe 1. Escolhemos essa variável como possível target pois pode haver interesse em prever a probabilidade de um funcionário deixar a companhia.

A variável 3 é binária e está balanceada no dataset onde aproximadamente 50% dos exemplos pertencem a cada classe. Optamos por essa variável como possível target pois pode haver interesse em prever a probabilidade de um funcionário estar disposto a se mudar.

Os problemas nas variáveis 2 e 3 são problemas a serem modelados como classificação, enquanto o problema da variável 1 adotamos o método de regressão.

3. MODELAGEM

Decidimos realizar primeiro a modelagem do problema de regressão sobre a variável **left_company**. Para realizar a modelagem utilizamos um Label Encoder para substituir as variáveis categóricas por variáveis numéricas e em alguns testes reduzimos as features referentes a pesquisas terceirizadas pela média das pesquisas para cada funcionário e então criamos um baseline utilizando uma Random Forest. Aplicando métricas de validação notamos que o modelo estava performando praticamente sem erro então houve a suspeita de overfitting, para verificar a interação das features com o modelo utilizamos da biblioteca Shap. Percebemos que o modelo estava extremamente dependente da variável **ID**, que não faz sentido algum para a previsão de avaliação então removemos esta variável e treinamos novamente a random forest no novo conjunto de dados.

Notamos agora uma forte dependência com a variável **percent_remote** que faz referência ao percentual de trabalho que o funcionário realiza de maneira remota, essa informação pode ser relevante mas a dependência era extremamente alta então decidimos remover a variável para verificar o comportamento do modelo. Com a remoção da variável as variáveis com maior poder preditivo passaram a ser o tempo médio (em horas) que o funcionário trabalha no mês, e a identificação do funcionário com a empresa e com seu cargo. O gráfico da relação de influência das variáveis neste modelo específico é mostrado abaixo.

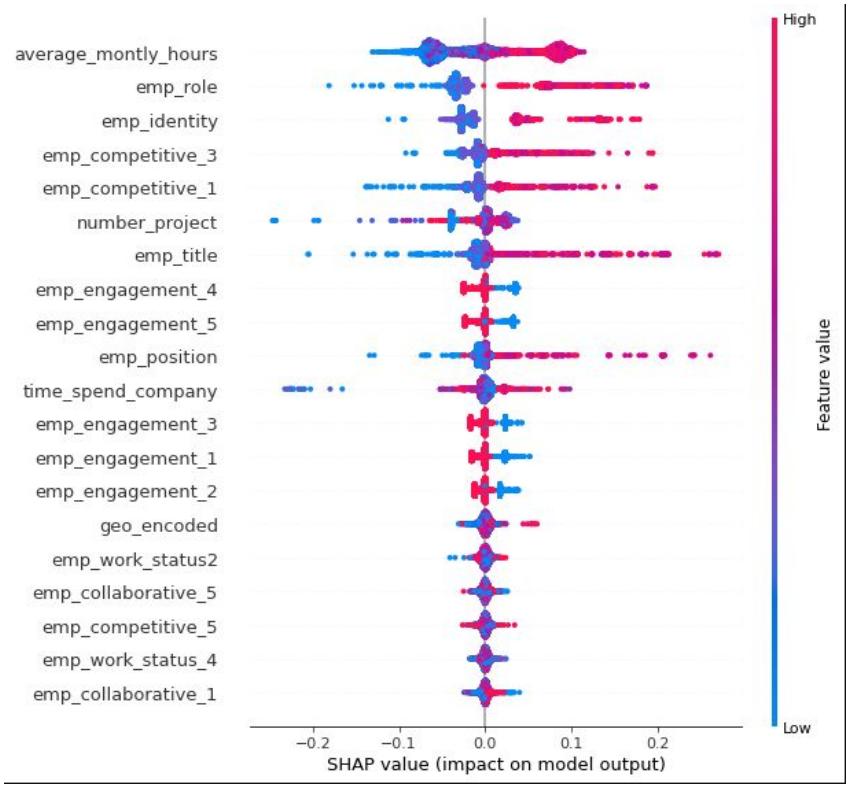

Figura 1

Analisando esse gráfico percebemos que o aumento de horas trabalhadas faz com que o modelo aumente o resultado da avaliação do funcionário. Além disso, quanto mais o funcionário se identifica com a empresa e com seu cargo também faz com que o modelo aumente o valor da avaliação do funcionário.

Além de criar um baseline utilizando Random Forest, também criamos uma baseline utilizando o Lightgbm. O Lightgbm é sensivel à overfiting porém decidimos utilizar este modelo pois o conjunto de dados, apesar de não ser grande, não é extremamente pequeno e esse modelo apresenta complexidade de tempo e de memória inferior à Random Forests devido à sua arquitetura.

As métricas utilizadas para analisar os dois modelos foram erro médio absoluto e erro médio quadrático. Os dois modelos “crus” tiveram resultados muito similares, então criamos um otimização bayesiana buscando hiperparâmetros que minimizam o erro médio absoluto do modelo lightbm porém o resultado dessa otimização foi mínimo.

Para finalizar o baseline aplicamos validação cruzada sobre a random forest para ter certeza que a métrica inicial estava generalista e a hipótese foi confirmada. Também testamos um ensemble básico com média simples e média ponderada sobre os dois modelos baseline para ver se influenciava significativamente no resultado final mas o resultado foi mínimo também.