Este projeto é parte do curso de data science e engenharia de dados da Stack Academy e tem como objetivo cobrir todas as etapas de um projeto real de Data Science.

Foram utilizados dados para responder questões da área de recursos humanos, tais como:

* Quais são os fatores que influenciam para um colaborador deixar a
* empresa?
* Como reter pessoas?
* Como antecipar e saber se um determinado colaborador vai sair da empresa?

Além destas questões, este projeto procurou disponibilizar recursos para que a empresa consiga realizar a predição para verificar se um colaborador vai ou não deixar a empresa com base em atributos como comportamento e carga de trabalho, nível de satisfação com a empresa e resultados de performance.

Para resolver esse problema foi construída uma solução completa para armazenamento, gestão e automatização de fluxos de dados utilizando tecnologias como Apache Airflow, Docker e Minio além de explorar uma suíte poderosa de tecnologias para trabalhar com Análise de Dados e Machine Learning que são: Pandas, Scikit-learn, Pycaret, SweetViz, Streamlit.

Depois da infraestrutura devidamente criada e configurada, levando em consideração o desafio proposto foram criados e modelados atributos relevantes para análise utilizando fontes de dados diversas como arquivos em formato xlsx, json e dados no Sistemas de Gerenciamento de Banco de Dados MySQL.

Na etapa de Análise Exploratória de Dados foram descobertos os vários insights importantes abaixo:

A empresa tem uma rotatividade (turnover) de 24%.

Podemos assumir que os empregados que mais deixam a empresa estão menos satisfeitos.

Existe um valor considerável de empregados insatisfeitos.

A maioria dos empregados que saíram tinha salário baixo ou médio.

Os departamentos de vendas, técnico e suporte são os 3 departamentos com maior índice de turnover.

Todos os colaboradores que estavam inseridos sem muitos projetos deixaram a empresa.

Colaboradores com baixa performance tendem a deixar a empresa.

Colaboradores insatisfeitos com a empresa têm uma maior tendência para evadir.

Por meio da análise foi possível desenvolver 3 grupos distintos para agrupar colaboradores que deixaram a empresa por comportamentos similares:

Grupo 1: Colaboradores insatisfeitos e trabalhadores

Grupo 2: Com baixa performance e insatisfeitos

Grupo 3: Empregados satisfeitos e trabalhadores

Para a estimativa com o objetivo de predizer se um empregado vai deixar a empresa foi implementado um modelo utilizando o algoritmo Gradient Boosting Classifier que atingiu uma performance de AUC em 0.80.

Neste projeto foi possível praticar e implementar conceitos importantes da ciência e engenharia de dados e propor uma solução para um problema da área de Recurso Humanos.

Como um processo de melhoria contínua podemos desenvolver uma automação para executar não só o pipeline de coleta e transformação de dados como automatizar os passos da etapa de Machine Learning e Deploy.