Asiana Holloway
October 6, 2024
STAT5165

Project Proposal: Analyzing Employee Attrition Using Spark and Python

Employee Attrition is an issue for many organizations as it can signal underlying problems within the workplace, such as low job satisfaction or inadequate support for employees. Understanding why employees leave can help companies identify areas for improvement and retain valuable talent. The goal of this project is to analyze the primary factors influencing employee turnover using a simple linear regression model. By leveraging Spark's powerful data processing capabilities, I aim to identify which variables (e.g., job satisfaction, monthly income, and work-life balance) are most strongly linked to employees leaving their jobs.

Some objectives I have for this project include using Spark and Python to process a real-world employee dataset. I want to Implement basic data preprocessing techniques, such as handling missing values and converting categorical data into a numeric format. I will also perform a linear regression analysis to identify which factors are significantly associated with employee Attrition. Once this process is complete, I will be ready to present my findings in a clear and concise manner, making the analysis accessible to both technical and non-technical stakeholders.

For this project, I will be using the IBM HR Analytics Employee Attrition Dataset, which contains over 1,470 records. This dataset includes various employee attributes, such as age, monthly income, job role, and job satisfaction, alongside the key indicator of whether the employee has left the company. Using this dataset, I will explore patterns and correlations that contribute to Attrition rates. A link to my data source can be viewed below: Dataset Source: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

This dataset is also available on my GitHub repository for easier access: GitHub Dataset Link To achieve the project goals, the implementation will follow a structured yet simplified approach. I will create a Spark cluster using two virtual machines (VMs). One will act as the master node while the other will serve as a worker node. This setup is important to demonstrate distributed computing. Next, I will load the dataset into Spark DataFrames and handle any missing values by either filling them with appropriate defaults or removing the affected rows. Finally, I will Transform categorical variables (e.g., Job Role, Department) into numeric values using label encoding, making them suitable for statistical analysis.

I chose Linear Regression Analysis due to its simplicity and it was the best fit for this project. The analysis will help measure the impact of variables like Age, Monthly Income, and Job Satisfaction on whether an employee leaves the organization. Then I will compute correlation coefficients to measure the strength of each factor's relationship with Attrition.

The results of the linear regression model will be summarized, highlighting the most significant predictors of Attrition. To make the findings more insightful, I will visualize the top factors influencing Attrition using basic bar charts or scatter plots.

The following tools and Python libraries will be used for the project: Spark, for distributed data processing and handling large datasets. **Pyspark**, Spark's Python library for managing Spark DataFrames and executing data transformations. Pandas, for local data

Asiana Holloway
October 6, 2024
STAT5165

Project Proposal: Analyzing Employee Attrition Using Spark and Python

manipulation and basic visualizations. Then scikit-learn, to implement linear regression and compute correlation statistics.

By the end of this project, I expect to provide a clear and concise demonstration of how to set up a Spark environment, preprocess data, and run a basic linear regression analysis. The insights gained from this analysis will help understand which factors are most strongly associated with employee Attrition. This will not only make the project understandable for newcomers to data science but also offer practical value for companies looking to address employee retention challenges.

The main challenges for this project involve setting up a functional Spark cluster using two machines and ensuring proper data handling during the preprocessing stage. However, these obstacles will be tackled by keeping the coding and setup instructions straightforward and simple.

This project serves as an introduction to working with Spark for data analysis in a distributed computing environment. It aims to provide a hands-on experience in a way that is approachable and straight forward. Through this project, I hope to simplify complex data analysis concepts and present actionable insights for reducing employee Attrition.