**Waiter tips prediction with machine learning**

**Problem statement:**

Our aim is to predict the tip amount that a waiter may receive, considering various factors such as the type of restaurant, the number of people dining, the total bill amount, and other relevant factors. As tipping behavior is influenced by multiple variables, our objective is to develop a model that can accurately forecast the expected tip amount based on these factors.

**Data collection and preprocessing:**

Data collection was carried out by searching for publicly available datasets on restaurant bills and tips. The dataset included various features such as the number of customers, total bill amount, and tip amount, as well as additional information such as whether the customer was a smoker or non-smoker, the day of the week, whether it was a dinner or lunch, and the customer's gender. However, upon initial inspection, I noticed some outliers and anomalies in the data. Therefore, I took necessary steps to preprocess the data, which included identifying and removing any invalid or missing data points and checking for any potential data skewness or outliers that may adversely impact the performance of our model.

**Feature engineering:**

For feature engineering, I used the StringIndexer method to convert categorical variables such as sex, time, smoker or non-smoker, and day of the week into numerical values. Next, I used the VectorAssembler method to combine these variables, along with the total bill column, into a single feature vector. This allowed to create a more comprehensive set of features that captured the relationships between different variables in the data. Finally, I split the data into training and testing sets using the randomSplit method, which allowed us to train and evaluate our machine learning model on separate datasets. Overall, these feature engineering techniques helped to extract more meaningful information from the data and improve the accuracy of our predictive model.

**Model selection:**

Linear regression was well-suited to our dataset as it allowed to model the relationship between the independent variables (such as total bill, day of the week, and customer gender) and the dependent variable (waiter's tip) with a linear equation. Additionally, it performed well on evaluation metric, such as root mean squared error indicating that it was able to accurately predict tips given the input features.

**Model training and evaluation:**

For the waiter tip prediction model, the data was split into 80% training and 20% validation sets. Linear regression was then applied on the training data and the performance was evaluated using the Root Mean Squared Error (RMSE) metric. The obtained RMSE value was 0.98, which indicates that the model has an average error of approximately $0.98 on tip prediction. This evaluation demonstrates that the linear regression model can be a suitable choice for predicting tip amounts for waiters.

**Results and insights:**

From the coefficients, the time, sex, and smoker variables have the greatest impact on the predicted tip amount, with larger absolute values than the other variables. Specifically, a unit increase in time is associated with a decrease in the predicted tip amount, while being a male (sex = 0) and being a smoker (smoker = 1) are associated with an increase in the predicted tip amount, all else being equal.

**Libraries used:**

Import pandas as pd

Import plotly.express as px

from pyspark.sql import SparkSession

from pyspark.sql.functions import \*

from pyspark.sql.types import \*

from pyspark.ml.feature import StringIndexer,VectorAssembler

from pyspark.ml import Pipeline

from pyspark.ml.regression import LinearRegression

from pyspark.ml.evaluation import RegressionEvaluator