```
In [11]: from pyspark.sql import SparkSession
         spark=SparkSession.builder.appName('ML model').getOrCreate()
```

```
In [17]: FEBRUARY\26 feb (time series and spark)\ml.csv",header=True,inferSchema=True)
```

```
In [20]: training.show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
|    Sunny| 29|         4| 20000|
|     Paul| 24|         3| 20000|
|   Harsha| 21|         1| 15000|
|  Shubham| 23|         2| 18000|
+---------+---+----------+------+
```

```
In [21]: training.printSchema()
```

```
root
 |-- Name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- Experience: integer (nullable = true)
 |-- Salary: integer (nullable = true)
```

```
In [22]: training.columns
```

```
Out[22]: ['Name', 'age', 'Experience', 'Salary']
```

[Age,Experience]----> new feature--->independent feature

```
In [26]: from pyspark.ml.feature import VectorAssembler
```

```
In [27]: featureassembler=VectorAssembler(inputCols=["age","Experience"],outputCol="Ind
```

```
In [32]: output=featureassembler.transform(training)
```

In [33]:
```
output.show()
```

```
+---------+---+----------+------+--------------------+
|     Name|age|Experience|Salary|Independent Features|
+---------+---+----------+------+--------------------+
|    Krish| 31|        10| 30000|         [31.0,10.0]|
|Sudhanshu| 30|         8| 25000|          [30.0,8.0]|
|    Sunny| 29|         4| 20000|          [29.0,4.0]|
|     Paul| 24|         3| 20000|          [24.0,3.0]|
|   Harsha| 21|         1| 15000|          [21.0,1.0]|
|  Shubham| 23|         2| 18000|          [23.0,2.0]|
+---------+---+----------+------+--------------------+
```

In [34]:
```
output.columns
```

Out[34]:
```
['Name', 'age', 'Experience', 'Salary', 'Independent Features']
```

In [35]:
```
finalized_data=output.select("Independent Features","Salary")
```

In [37]:
```
finalized_data.show()
```

```
+--------------------+------+
|Independent Features|Salary|
+--------------------+------+
|         [31.0,10.0]| 30000|
|          [30.0,8.0]| 25000|
|          [29.0,4.0]| 20000|
|          [24.0,3.0]| 20000|
|          [21.0,1.0]| 15000|
|          [23.0,2.0]| 18000|
+--------------------+------+
```

In [38]:
```
from pyspark.ml.regression import LinearRegression
##train test split
train_data,test_data=finalized_data.randomSplit([0.75,0.25])
regressor=LinearRegression(featuresCol='Independent Features', labelCol='Salar
regressor=regressor.fit(train_data)
```

In [39]:
```
regressor.coefficients
```

Out[39]:
```
DenseVector([-518.2482, 2094.8905])
```

In [40]:
```

regressor.intercept
```

Out[40]:
```
24605.839416054776
```

In [41]:

```
pred_results=regressor.evaluate(test_data)
```

In [44]:

```
pred_results.predictions.show()
```

```
+--------------------+------+------------------+
|Independent Features|Salary|        prediction|
+--------------------+------+------------------+
|          [24.0,3.0]| 20000| 18452.55474452559|
|          [29.0,4.0]| 20000|17956.204379562776|
+--------------------+------+------------------+
```

In [45]:

```
pred_results.meanAbsoluteError,pred_results.meanSquaredError
```

Out[45]:

```
(1795.6204379558167, 3285843.6784043186)
```

In [ ]:

```

```