STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

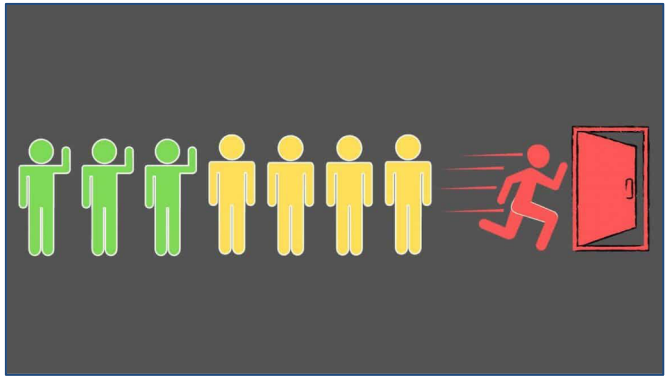
Telecom Churn Prediction

BIA 678: Big Data Technologies
Team 10

Pranay Bhandare
Harman Singh Bath
Jungho Park

Project Purpose

- *Customer retention* is a huge problem in the telecommunication industry and can prove to be less costly than attracting a new one
    - There are data points on customers that can be analyzed to strategize on how to retain them when realized they are unhappy with a company's services
- Businesses want to maximize their number of customers. To achieve this goal, it is equally important to attract new ones but also retain existing customers
- Building up and keeping a loyal clientele can be challenging, especially when customers are free to choose from a variety of telecommunication providers and their diverse products/services
- To compare the performance of different classification models: *Logistic Regression, Decision Tree, Linear Support Vector Machine (Linear SVM)*
- Analysis on the impact of *scaling* using AWS clusters



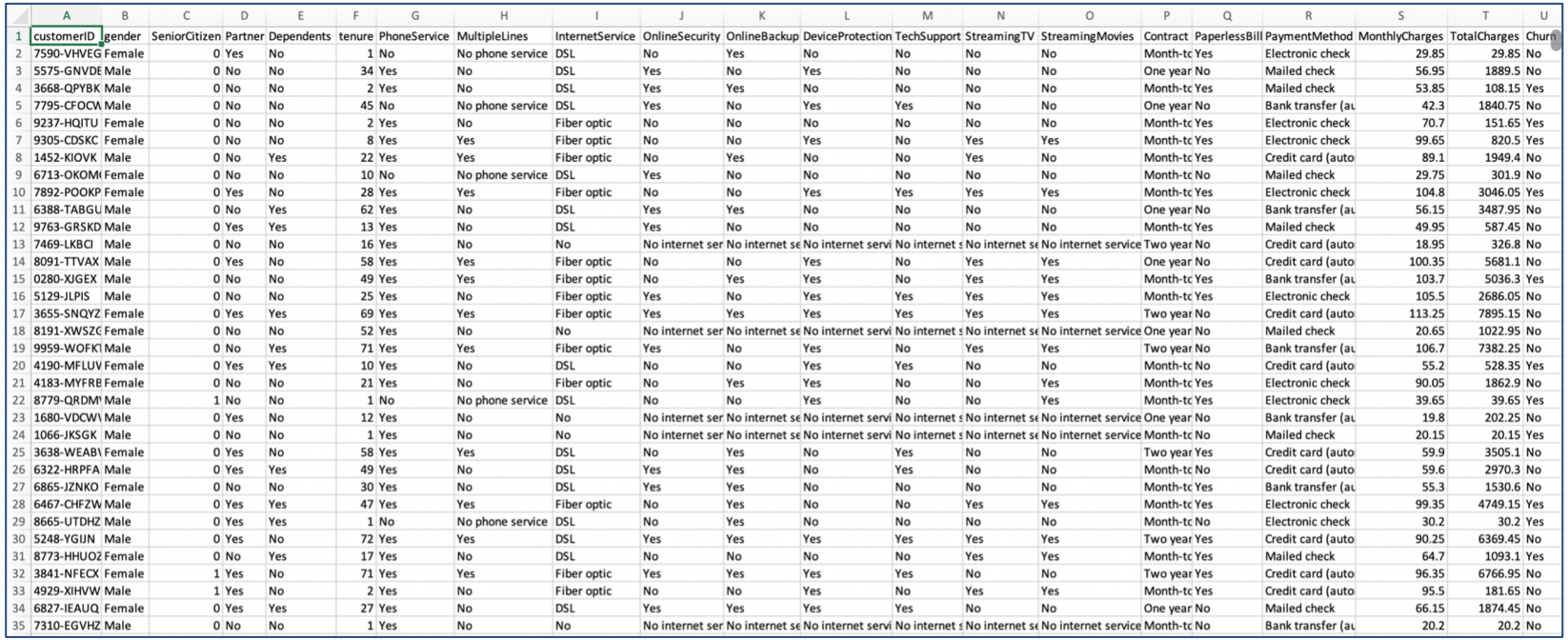
Data Description

- The dataset found on Kaggle contains information about approximately *6,000 users, their demographic characteristics, the services they use, the duration of using the company's services, payment methods and information*

- Our main task is to analyze this dataset and predict the user churn rate
  - To further identify people who will and will not renew their contract

```
|-- _c0: integer (nullable = true)
|-- customerID: string (nullable = true)
|-- gender: string (nullable = true)
|-- SeniorCitizen: integer (nullable = true)
|-- Partner: string (nullable = true)
|-- Dependents: string (nullable = true)
|-- tenure: integer (nullable = true)
|-- PhoneService: string (nullable = true)
|-- MultipleLines: string (nullable = true)
|-- InternetService: string (nullable = true)
|-- OnlineSecurity: string (nullable = true)
|-- OnlineBackup: string (nullable = true)
|-- DeviceProtection: string (nullable = true)
|-- TechSupport: string (nullable = true)
|-- StreamingTV: string (nullable = true)
|-- StreamingMovies: string (nullable = true)
|-- Contract: string (nullable = true)
|-- PaperlessBilling: string (nullable = true)
|-- PaymentMethod: string (nullable = true)
|-- MonthlyCharges: double (nullable = true)
|-- TotalCharges: string (nullable = true)
|-- Churn: string (nullable = true)
```

Data Description (cont.)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBill	PaymentMethod	MonthlyCharges	TotalCharges	Churn
2	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-to	Yes	Electronic check	29.85	29.85	No
3	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
4	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to	Yes	Mailed check	53.85	108.15	Yes
5	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank transfer (au	42.3	1840.75	No
6	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to	Yes	Electronic check	70.7	151.65	Yes
7	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to	Yes	Electronic check	99.65	820.5	Yes
8	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to	Yes	Credit card (auto	89.1	1949.4	No
9	6713-OKOM	Female	0	No	No	10	No	No phone service	DSL	Yes	No	No	No	No	No	Month-to	No	Mailed check	29.75	301.9	No
10	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to	Yes	Electronic check	104.8	3046.05	Yes
11	6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank transfer (au	56.15	3487.95	No
12	9763-GRSKD	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to	Yes	Mailed check	49.95	587.45	No
13	7469-LKBCI	Male	0	No	No	16	Yes	No	No	No internet ser	No internet se	No internet servi	No internet s	No internet se	No internet service	Two year	No	Credit card (auto	18.95	326.8	No
14	8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	One year	No	Credit card (auto	100.35	5681.1	No
15	0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to	Yes	Bank transfer (au	103.7	5036.3	Yes
16	5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Month-to	Yes	Electronic check	105.5	2686.05	No
17	3655-SNQYZ	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit card (auto	113.25	7895.15	No
18	8191-XWSZG	Female	0	No	No	52	Yes	No	No	No internet ser	No internet se	No internet servi	No internet s	No internet se	No internet service	One year	No	Mailed check	20.65	1022.95	No
19	9959-WOFK	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank transfer (au	106.7	7382.25	No
20	4190-MFLUV	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to	No	Credit card (auto	55.2	528.35	Yes
21	4183-MYFRB	Female	0	No	No	21	Yes	No	Fiber optic	No	Yes	Yes	No	No	Yes	Month-to	Yes	Electronic check	90.05	1862.9	No
22	8779-QRDM	Male	1	No	No	1	No	No phone service	DSL	No	No	Yes	No	No	Yes	Month-to	Yes	Electronic check	39.65	39.65	Yes
23	1680-VDCW	Male	0	Yes	No	12	Yes	No	No	No internet ser	No internet se	No internet servi	No internet s	No internet se	No internet service	One year	No	Bank transfer (au	19.8	202.25	No
24	1066-JKSGK	Male	0	No	No	1	Yes	No	No	No internet ser	No internet se	No internet servi	No internet s	No internet se	No internet service	Month-to	No	Mailed check	20.15	20.15	Yes
25	3638-WEAB	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	Yes	No	No	Two year	Yes	Credit card (auto	59.9	3505.1	No
26	6322-HRPFA	Male	0	Yes	Yes	49	Yes	No	DSL	Yes	Yes	No	Yes	No	No	Month-to	No	Credit card (auto	59.6	2970.3	No
27	6865-JZNKO	Female	0	No	No	30	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to	Yes	Bank transfer (au	55.3	1530.6	No
28	6467-CHFZW	Male	0	Yes	Yes	47	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	Yes	Month-to	Yes	Electronic check	99.35	4749.15	Yes
29	8665-UTDHZ	Male	0	Yes	Yes	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-to	No	Electronic check	30.2	30.2	Yes
30	5248-YGIJN	Male	0	Yes	No	72	Yes	Yes	DSL	Yes	Yes	Yes	Yes	Yes	Yes	Two year	Yes	Credit card (auto	90.25	6369.45	No
31	8773-HHUOZ	Female	0	No	Yes	17	Yes	No	DSL	No	No	No	No	Yes	Yes	Month-to	Yes	Mailed check	64.7	1093.1	Yes
32	3841-NFECX	Female	1	Yes	No	71	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	No	No	Two year	Yes	Credit card (auto	96.35	6766.95	No
33	4929-XIHVW	Male	1	Yes	No	2	Yes	No	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to	Yes	Credit card (auto	95.5	181.65	No
34	6827-IEAUQ	Female	0	Yes	Yes	27	Yes	No	DSL	Yes	Yes	Yes	Yes	No	No	One year	No	Mailed check	66.15	1874.45	No
35	7310-EGVHZ	Male	0	No	No	1	Yes	No	No	No internet ser	No internet se	No internet servi	No internet s	No internet se	No internet service	Month-to	No	Bank transfer (au	20.2	20.2	No

Data Description (cont.)

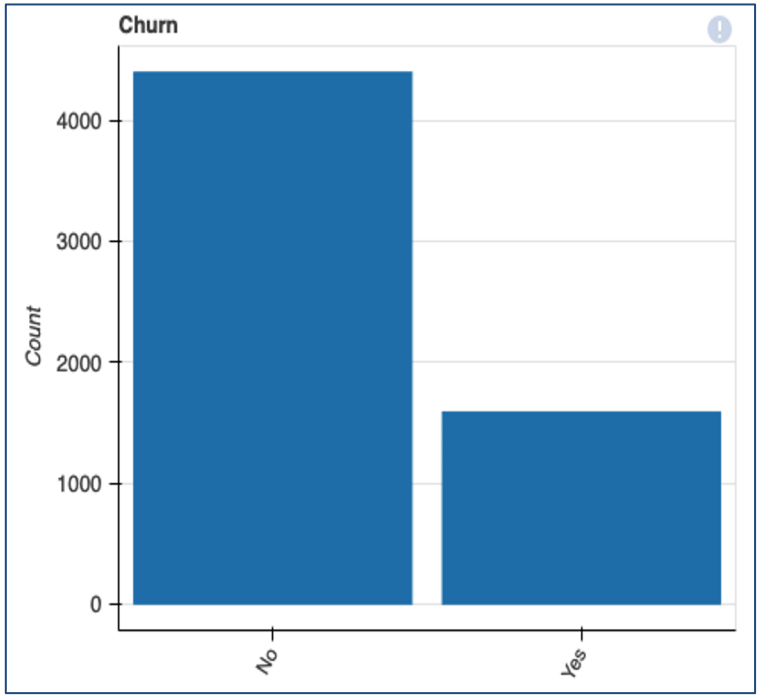
Dataset Statistics

Number of Variables	21
Number of Rows	5986
Missing Cells	0
Missing Cells (%)	0.0%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	6.7 MB
Average Row Size in Memory	1.1 KB
Variable Types	Categorical: 18 Numerical: 3

Dataset Insights

`SeniorCitizen` is skewed	Skewed
`tenure` is skewed	Skewed
`MonthlyCharges` is skewed	Skewed
`customerID` has a high cardinality: 5986 distinct values	High Cardinality
`TotalCharges` has a high cardinality: 5611 distinct values	High Cardinality
`customerID` has constant length 10	Constant Length
`customerID` has all distinct values	Unique
`SeniorCitizen` has 5020 (83.86%) zeros	Zeros

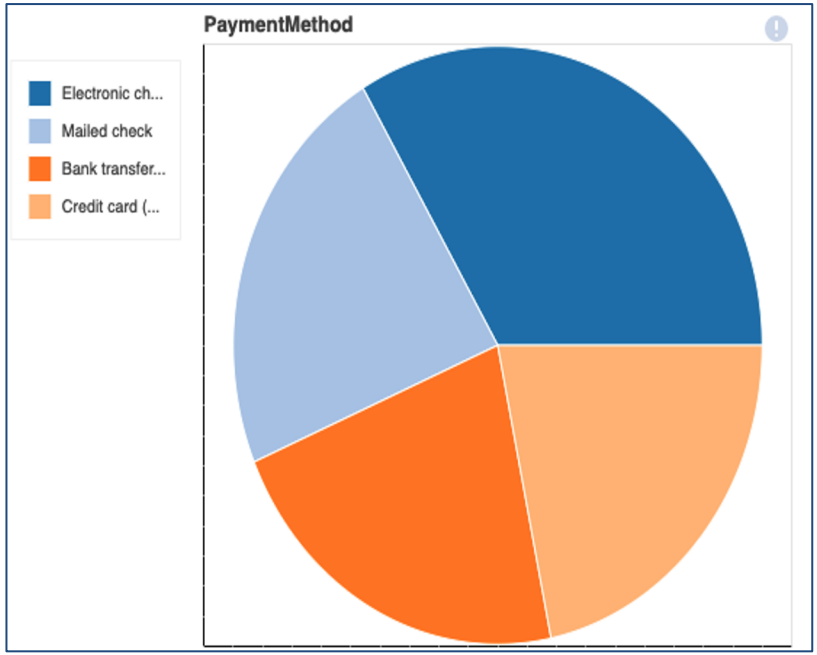
Exploratory Data Analysis (Churn)

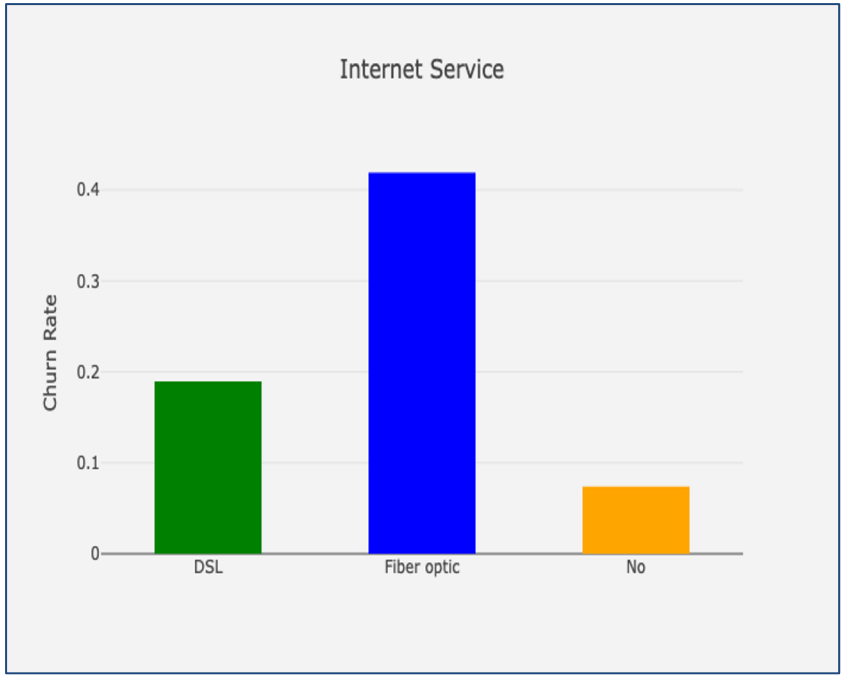



```
df_data.groupby('gender').Churn.mean()
```

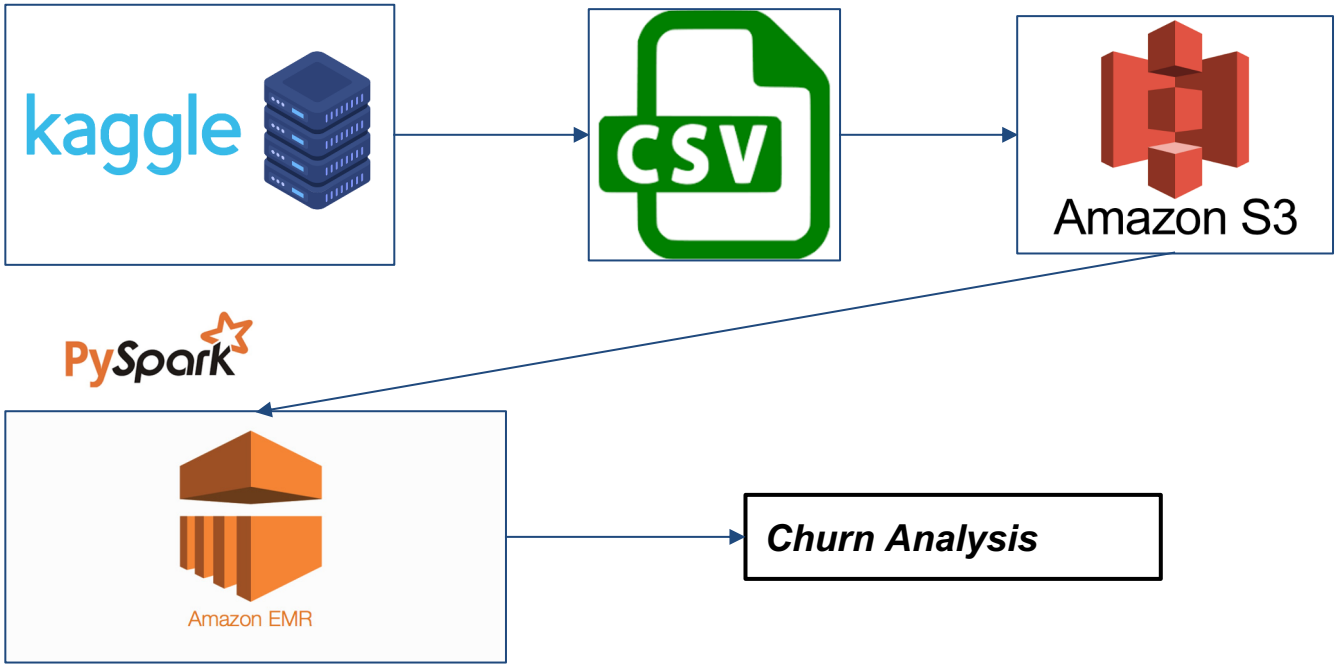
```
gender
Female    0.269209
Male      0.261603
Name: Churn, dtype: float64
```

Exploratory Data Analysis (Cont.)





Data Pipeline



Data Preprocessing

- Built a pipeline using Pyspark
- We first loaded the data and put all the numerical features into a vectorAssembler
- For the categorical features, we took them through a stringIndexer and performed one-hot encoding
- Create a new column ‘features’ which displays a vectorized list of binary values
- Last, we fit the pipeline on the training split and transform both the training and test splits (70/30 random split)

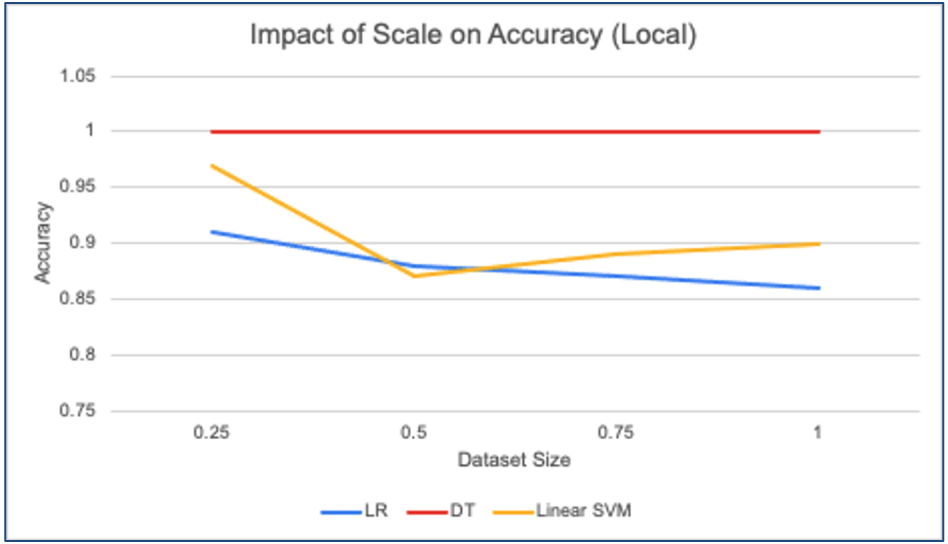
	label	features	...
0	0.0	(1.0, 0.0, 0.0, 1.0, 0.0, 1.0, 0.0, 0.0, 0.0, ...	...
1	0.0	(0.0, 1.0, 1.0, 1.0, 1.0, 0.0, 1.0, 0.0, 1.0, ...	...
2	1.0	(0.0, 0.0, 1.0, 1.0, 0.0, 1.0, 1.0, 0.0, 1.0, ...	...
3	0.0	(1.0, 1.0, 1.0, 1.0, 1.0, 0.0, 0.0, 1.0, 1.0, ...	...
4	0.0	(1.0, 1.0, 1.0, 1.0, 1.0, 0.0, 0.0, 1.0, 0.0, ...	...

[5 rows x 22 columns]

Impact of Scale on Quality of Analysis (Local)

Conclusions

- Unusual that the DT model had 100% accuracy for all dataset sizes
- Highest accuracy is achieved with least dataset size
- Contradicts belief that more data results in higher accuracy scores

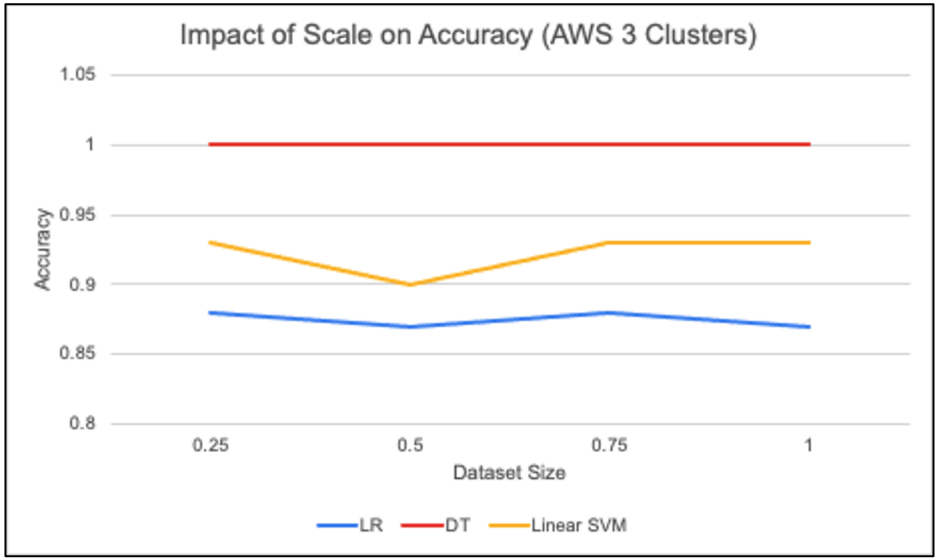


Impact of Scale on Quality of Analysis (AWS)

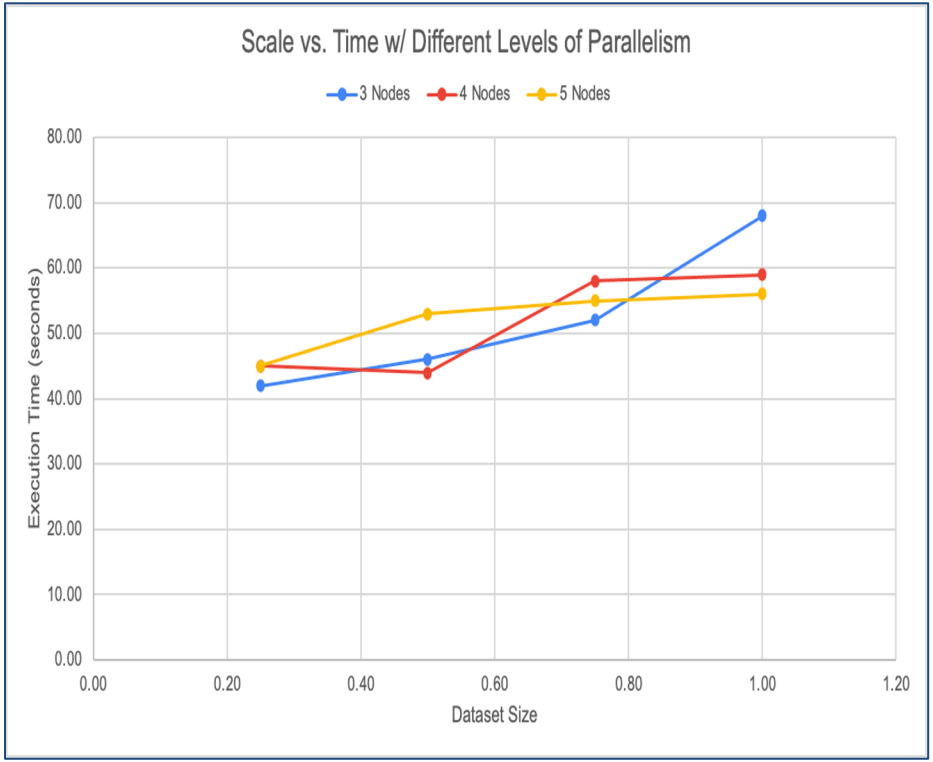
Conclusions

- Accuracy values are from AWS cluster size of 3 (1 master, 2 slave nodes)
- Not a clear relationship between dataset size and accuracy
- Unusual that the DT model had 100% accuracy for all dataset sizes
- DT model is not reliable at this scale. We need more data to validate.
- Quality of the data has more significant impact on accuracy for classification problems than scale (dataset size)

```
[[1327    0]
 [   0  480]]
1.0
```



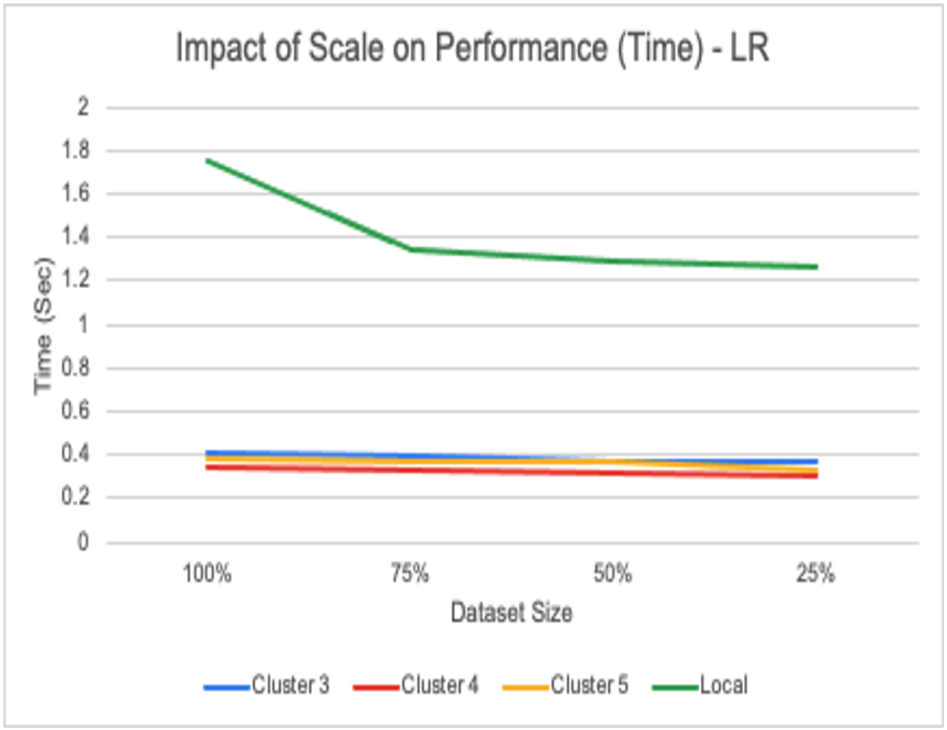
Impact of Scale on Performance (Time)



Conclusions

1. As the dataset size increases, there is a general increase in execution time
2. Inverse relationship between number of nodes and execution time

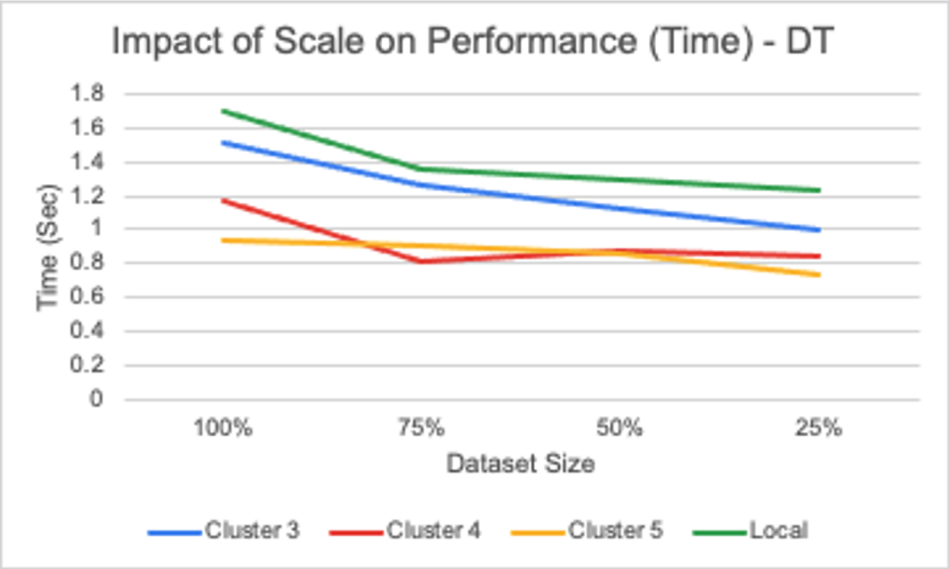
Impact of Parallel Computation on Performance With Respect to Scalability



- Local environment takes longer time than AWS parallel environment.
- Gradual decrease in time as dataset size decreases

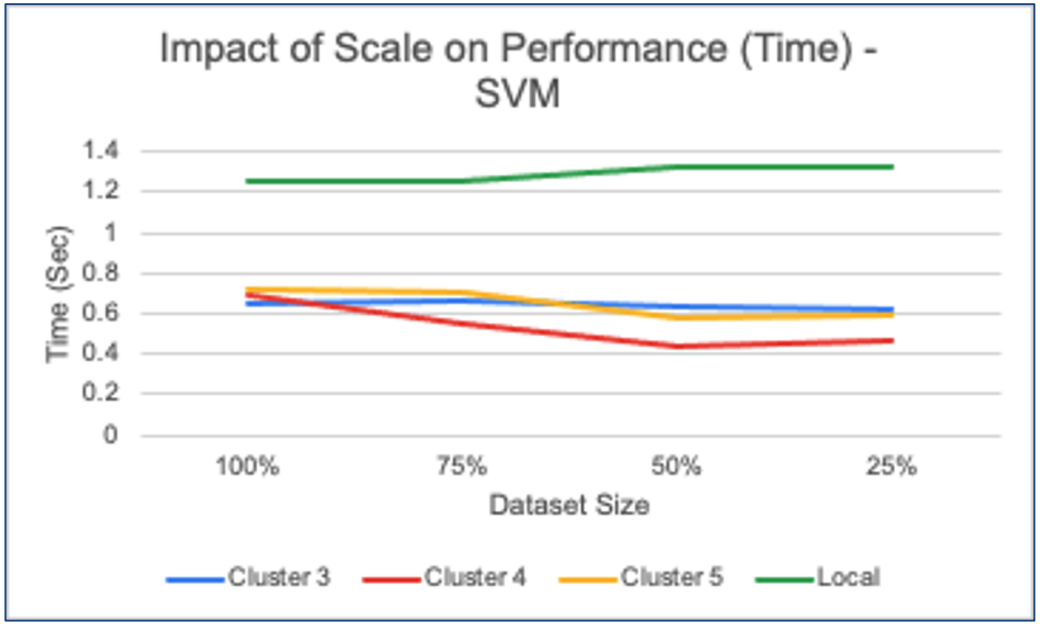
Execution Time (LR)	Cluster Size			
Dataset Size	Cluster 3	Cluster 4	Cluster 5	Local
100%	0.411	0.349	0.383	1.75
75%	0.391	0.324	0.365	1.34
50%	0.375	0.314	0.365	1.29
25%	0.372	0.305	0.333	1.26

Impact of Parallel Computation on Performance With Respect to Scalability (cont.)



- Local environment proves to take the most time
- Gradual decrease in time as dataset size decreases

Impact of Parallel Computation on Performance With Respect to Scalability (cont.)



- Local environment takes longer time than AWS.
- Decrease in dataset size does not necessarily decrease the time used.

Model Comparison Summary

1. Local non-parallel environment consumes more time than AWS parallel computing environments in every ML model.
2. As cluster size increases, execution time generally decreased
3. Decision Tree resulted in 100% accuracy
    a. Maybe due to over fitting
    b. Clean & simple dataset from Kaggle
    c. Not enough data
4. Percentage scaling matters
5. The dataset was really clean, but too small to conclude the impact of accuracy

Conclusion

Major Outcome:

- *AWS parallelization has a significant impact on time performance compared to local execution*
  - Difference is miniscule but evident through AWS nodes and cluster performance
  - Binary classification is a simple classification problem so scale and accuracy didn't prove an evident relationship
- *Time VS Scaling Plot tend to be flattened*
  - Assumptions: Current Dataset is not big enough to reveal the effect of Scaling on Run Time

Future Approach:

- Enhance more features to add complexity to the problem
- Collect more numerical features

THANK YOU!

QUESTIONS?