**Data source & About the data**

The dataset for this report comes from different data sources.

**Meta data-**

Twitter dataset acts as primary dataset, Which is an Text dataset

Text analysis

Positivity, Negativity, Intent…

Survey dataset acts as secondary dataset, which is an Multivariate dataset

Mostly it was built on Ordinal/Nominal/Categorical question time.

Third one is support dataset for our project,

where we supplied with all the historical Election details of 2011 and 2016.

Features descriptions

Twitter –

User, Location, Tweets

Survey –

General details – Age.

Gender.

whether pollster resisted to vote.

If not reason behind that. Election details – whether pollster understood the election process.

Whether they’ve Past voting experiences.

Did they contributed or collected money.

Basis of selecting their candidate.

Party details – Present CM ratings

Their experience towards ongoing government.

Few descriptive fields to diagnose the text analysis.

Finally, our Target field Expected CM/Party, 2021 in Election.

Historical –

District-wise assembly Constituency list and standing candidate.

Past election result were included.

**Data Cleaning**

Twitter data clean-up

Removing Duplicate (manual)

Stop words (will)

Emoji’s

Hyper links

Digits

Punctuations and

Non-English words

using all possible UNICODE values and reg ex.

Survey outliers – Appling domain knowledge instead of going for blind statistical methods.

Outliers 1 –

Pollster Registered to vote ? (No)

Voted in 2021 elections? (Yes)

Outliers 2 –

On the basis of selecting their political candidate

Party they’re supporting for.

Outliers 3 –

Voted in local election? (Yes)

Reason behind not voting in present election.

Outliers 4 –

On the basis of selecting their political candidate

Reasons behind voting in 2021

Outliers 5 –

Current CM Rating

Whether current government going in the Right directions.

**Deployment**

Initially we planned to host a Website, after checking the feasibility we found a Dashboard is quite enough and ended up in Tableau public.

**Limitations**

Sample size is not representing population size and hence lack of data acquisitions in survey data.

Survey data doesn’t represent entire constituencies.

Lack of Continuous numeric variables in our dataset.

Issues were faced on Twitter clean-up is not just countable.

Tableau limitation on plotting geo-graph District-wise instead of constituency.

Integrating Three different data source with different format.

Multi-class Target variables affected the Accuracy score.