A REPORT ON

ELECTION RESULT ANALYSIS

BY

Name of the Student	ID No.	Discipline
Achleshwar Luthra	2018A3PS0401P	B.E. EEE
Harsh Tripathi	2018B4AA0177G	M.Sc. Math +B.E. ECE
Harsh Yadav	2018A7PS0217P	B.E. CSE
Ishita Chakravarthy	2018B4AA0670G	M.Sc. Math +B.E. ECE
Kalash Shah	2018A7PS0213P	B.E. CSE
Nupur Funkwal	2018A7PS0624G	B.E. CSE
Shivansh Rustagi	2018A8PS0745P	B.E. ENI
Vijay Kumar Malhotra	2018B4AA0039H	M.Sc. Math +B.E. ECE

Prepared in fulfilment of the Practice School-I Course

AT

PASS Consulting, Hyderabad

A Practice School I station of

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

(May 2020)

A REPORT ON

ELECTION RESULT ANALYSIS

BY

Name of the Student	ID No.	Discipline
Achleshwar Luthra	2018A3PS0401P	B.E. EEE
Harsh Tripathi	2018B4AA0177G	M.Sc. Math +B.E. ECE
Harsh Yadav	2018A7PS0217P	B.E. CSE
Ishita Chakravarthy	2018B4AA0670G	M.Sc. Math +B.E. ECE
Kalash Shah	2018A7PS0213P	B.E. CSE
Nupur Funkwal	2018A7PS0624G	B.E. CSE
Shivansh Rustagi	2018A8PS0745P	B.E. ENI
Vijay Kumar Malhotra	2018B4AA0039H	M.Sc. Math +B.E. ECE

Prepared in fulfilment of the Practice School-I Course

AT

Pass Consulting, Hyderabad

A Practice School I station of

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

(May 2020)

ACKNOWLEDGEMENTS

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of the project. I am thankful for their aspiring guidance, invaluably constructive criticism, and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

I express my warm regards to Nishikanth V S for his support and guidance at Pass Consulting, Hyderabad. I would also like to thank my Practice School Faculty Prof. Shekhar Rajagopalan for his support.

ABSTRACT SHEET

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI (RAJASTHAN)

Practice School Division

Station: Pass Consulting **Centre:** Hyderabad

Duration: 6 weeks **Date of Start:** 18 May, 2020

Date of Submission: 27 June, 2020

Title of the Project: Election Result Analysis

ID No./Name/Discipline of the Student:

2018A7PS0624G

Nupur Funkwal

B. E. Computer Science

Name & Designations of the Experts: Nishikanth V S

Name of the PS Faculty: Professor Shekhar Rajagopalan

Key Words: Election Result Analysis, Data Analysis, Data Visualization, Political Science, Machine Learning

Project Areas: Political Consulting, Data Analysis, Machine Learning

Abstract:

The following report focuses on the progress of the project on predictive analysis of election results in India pursued as a part of Practice School - I internship at PASS Consulting, Hyderabad. The model aims to predict the winning party and the votes secured by the party in Assembly elections. It is built by analyzing data on the political and personal background of contesting candidates. The next step involved assigning weights to these parameters. A regression model was built which then underwent extensive training and fine tuning to achieve the best possible results.

Future prospects for the project include the clustering of various constituencies based on the similarities between their demographic parameters and other political factors.

Signature of Student

Signature of PS Faculty

Date: 27/06/2020

Date: 27/06/2020

Table of Contents

Introduction

This report focuses on the project of developing a predictive model for Legislative Assembly election results in India. It facilitates PASS consulting in predicting the election results on a constituency level, taking into account the political background of contesting candidates and past results. This report gives an understanding of how some of the quantifiable factors change in the process of election which can be later used for developing the model.

Company Overview

The PASS Consulting Group is an international group of companies with more than 700 employees. PASS was founded in 1981 and is based in Aschaffenburg, Germany. Their core competencies are IT-Consulting, Software-Development, Project-Management and Solution Providing.

The company consults with customers on how to optimize their business processes and systems. Their staff combines IT with expertise and they live up to their profession when it comes to consulting and executing complex IT projects. The PASS Research & Development team reviews and analyses trends, evaluates technologies and assesses them within the context of scientific studies.

The company's vision is the development of IT products, solutions and services in zero-effects quality and for market leadership in automated software generation. It is a quality enterprise that achieves perfection in customer orientation, organization and leadership, based on competences and values of truthfulness, responsibility and many others.

Organization Structure

We work under direct supervision of the director V Sudarshan of PASS GCA Consulting Pvt Ltd India. We have been in constant contact and support with the members of his team headed by V S Nishikanth.

Project

Objective of the project

This project aims to develop a data analysis model to forecast election results using the research outcomes of a study linking past electoral outcomes of various constituencies to their demographic parameters. It facilitates PASS consulting to predict the results of Assembly elections of a constituency at an assembly-constituency level. Another expected use case is to aid political parties in effectively choosing between prospective candidates depending on the input parameters of the candidate.

The Technologies

1. MS Excel
2. Python- Libraries used include Pandas, NumPy, Matplotlib, Seaborn, Plotly, Sklearn.
3. R- Libraries used include ggplot2, e1071, tidyverse and caret.
4. Algorithms used for the project are Linear and Multiple Regression, Decision trees and Random Forests.

Solution architecture

The first step was to collect data regarding the constituency demographics, infrastructure, previous results. The sources for this data had to be accurate and from government documents or from open source websites. After a lot of due research, SHRUG repository was chosen as an appropriate and reliable source for collecting this data, along with government records such as those of the ECI.

This data was then converted into CSV format which would help in further analysis and work as an input data to train our predictive model.

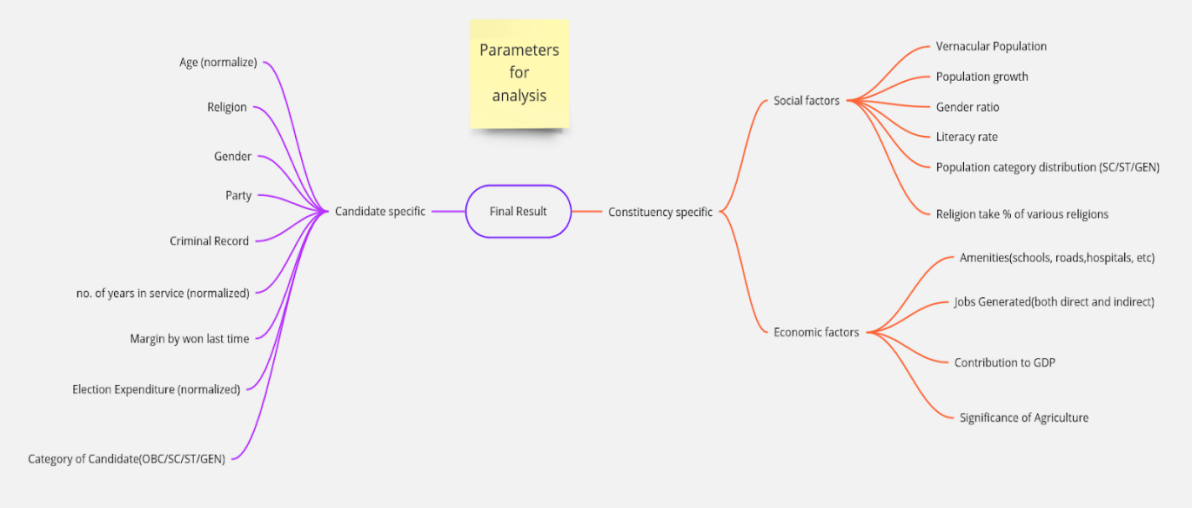
The data collected was raw and correlated to each other. So the data is now combined to form relationships to give us the raw structure of the data that would help in visualizing data better and get more information out of it.

The next step was to visualize the data using various libraries in Python and R. Scatter plots, clustered bar-charts and histograms gave us a good insight on the correlation of various parameters and also led us to some results which helped in

picking factors for the model and also verifying the weights assigned by the model to the various parameters.

The identification of candidate specific parameters led us to developing a preliminary model. This preliminary model was given the previous years' election data as input using python code. The labelEncoder function in the sklearn library was used to quantify the categorical variables. The multilinear regression analysis was done for each constituency and the weights obtained were analyzed and verified using various visualizations techniques. All the constituencies were divided into 3 groups based on the level of urbanization, the first group consisted of Bangalore south, Secunderabad and Hyderabad. The second group consisted of Nizamabad, Bellary and Chittor ,while Wayanad and Sivaganga constituted the third group. The weights obtained from the individual regression were then analyzed and averaged out to get a final set of weights for the three groups.

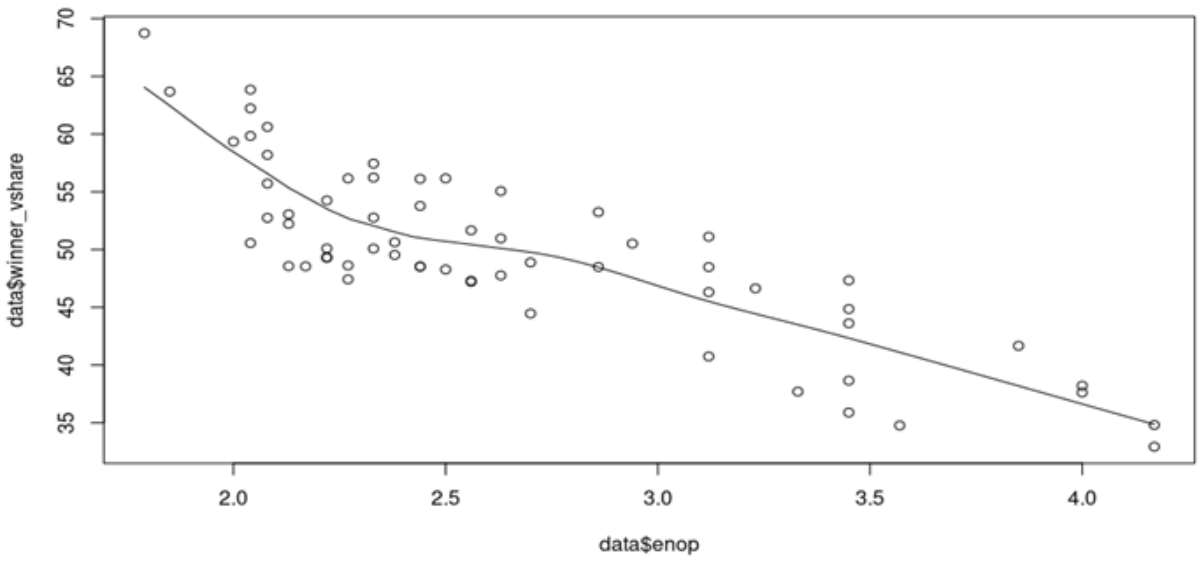
The final weights were then used to build a model for a predictive analysis, few algorithms such as K- Nearest Neighbours and Decision trees were also considered for the model but they had certain drawbacks that resulted in highly skewed results. Hence, Regression was identified to be the go to algorithm for the prediction model. Afterwards, the data was split into the test and training data. Using this, the regression algorithm was used to predict the winner candidate's vote share percentage. The model was then fine tuned and tested using the data from other constituencies based on whether the constituency was rural, semi-rural or urban.



(Figure showing the various parameters considered in the analysis)

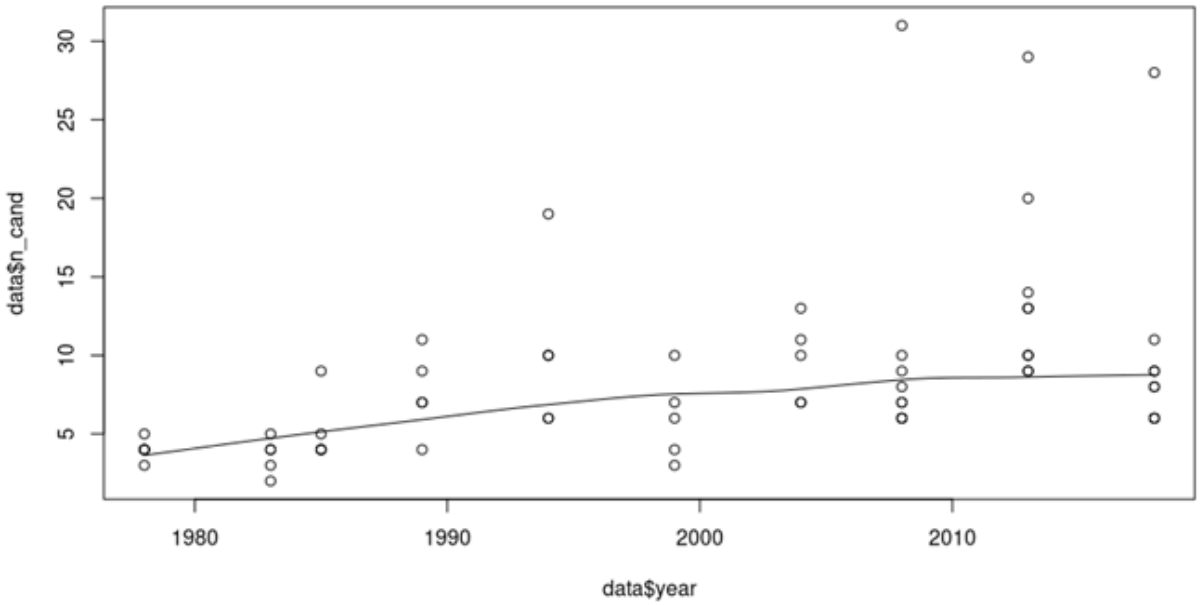
```
import numpy as np
import pandas as pd
import sklearn

from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn import preprocessing

df = pd.read_excel('/Users/pf/Desktop/test5.xlsx')
le = preprocessing.LabelEncoder()
df['tr_ac_name'] = le.fit_transform(df['tr_ac_name'])
df['party'] = le.fit_transform(df['party'])
df['deposit_lost'] = le.fit_transform(df['deposit_lost'])

cols = ['tr_ac_name','n_cand','turnout_percentage','enop','party','deposit_lost']
X = df[cols]
y = df['vote_share_percentage']
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.2, random_state = 10)
reg = LinearRegression()
reg.fit(X_train, y_train)
y_pred = reg.predict(X_test)
result_df = pd.DataFrame({'Actual' : y_test, 'Predicted' : y_pred})
print(result_df)
print(reg.score(X_test, y_test))
```

(The dataframe used in the above code snippet is with respect to the Legislative Assembly segments of Secunderabad)



(The scatter plot shows a strong correlation between the winning candidate's vote share and the effective no. of parties (enop*), data on assembly elections from 1978 to 2018 of all the assembly constituencies lying in the Bellary district constitute the data set.)

*enop = $N = \frac{1}{\sum_{i=1}^{n} p_i^2}$, n is the number of parties with at least one vote/seat and p is the proportion of all votes or seats won by the party.



(The scatter plot shows a monotonic rise in the number of contesting candidates over the last few decades, data on assembly elections from 1978 to 2018 of all the assembly constituencies lying in the Bellary district constitute the data set.)

Project description

This project focuses on building a model to predict the results of Assembly elections. For the primary analysis, 8 constituencies in South India were selected- Bangalore South, Bellari, Chittoor, Hyderabad, Nizamabad, Secunderabad, Sivaganga, and Wayanad. By analyzing these constituencies and collecting data regarding their demographics and previous election results, the goal was to identify various key parameters which influenced the elections. After identifying these parameters for various groups of constituencies, a predictive model had to be built. This predictive model had to provide information about the winning party and the percentage of votes secured by them in the elections.

I worked on Secunderabad constituency. Secunderabad Lok Sabha constituency comprises the following Legislative Assembly Segments.

- Mushreedabad
- Amberpet
- Khairatabad
- Jubilee Hills
- Sanathnagar
- Nampally
- Secunderabad

Data Collection

The first step to building a predictive model was to collect necessary data. Data regarding the candidates who have contested for elections, demographics of the constituency, infrastructure and various other factors which might play a role in the elections was collected and analysed.

Initially, demographic data was collected from Census 2011 and electoral data was collected using ECI website and MyNeta. Shrug Repository was also used for this purpose as it had both demographic & electoral data cleaned and formatted

properly in csv. This, however, led to a problem where Shrug doesn't map demographics data to constituencies which have overlapping borders with multiple district/ mandal/ villages (mostly for Urban Constituencies) [Shrug Codebook: pg 12,last para]. Hence, the demographic data for Secunderabad was missing completely. The Electoral data of the Assembly segments was obtained from Shrug itself.

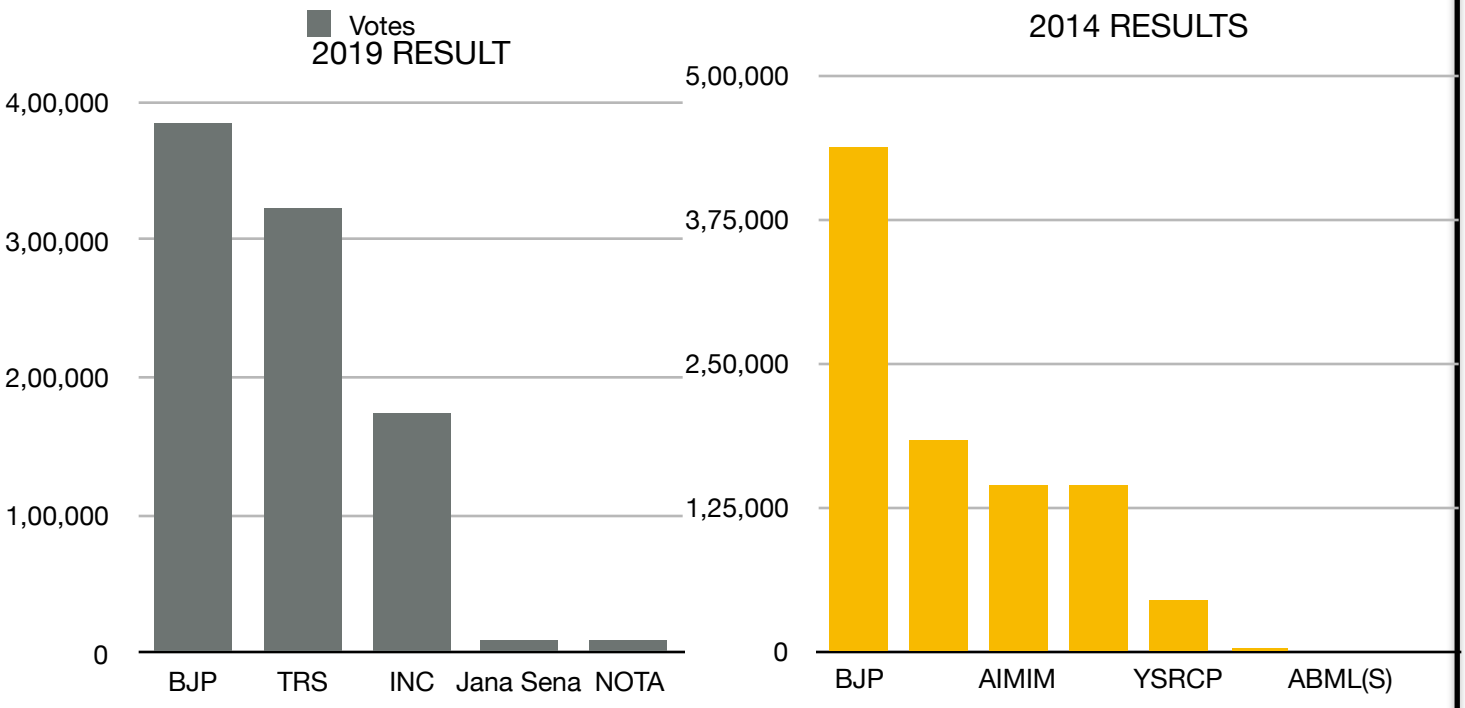
A major problem faced was, for urban constituencies, the census divisions (geographic) and the constituencies were not the same, resulting in no or poor mapping of the census data to constituency (one of the example). One workaround was approximating the subdistrict and clubbing their data together for using in Lok Sabha constituency election. Downside to this was the inaccuracy that can arise due to accumulation of a lot of data points and loss of useful micro-information (ward/subdistrict wise info) rendering this method unusable for Assembly Constituencies.

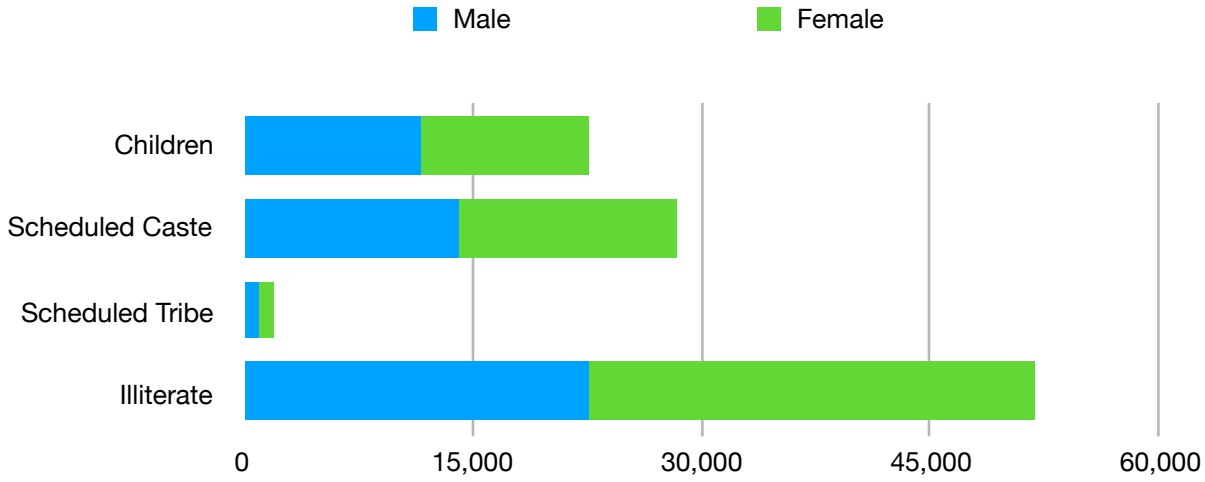
Moreover, though the data regarding other parameters was available, it was often not in an appropriate format to club it with the election result data. The data was present in the form of pdfs and images of old government documents. Using various online OCR tools**,** the data was converted into CSV format.

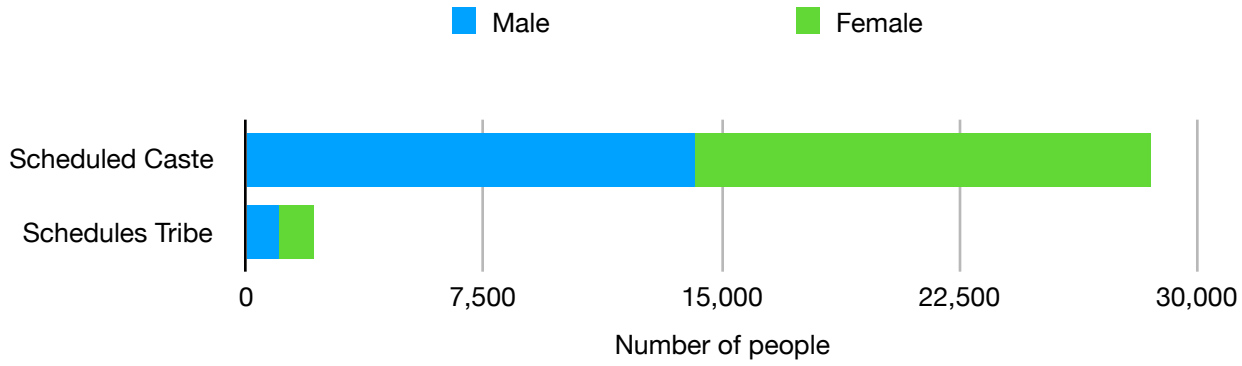
Data visualization and analysis

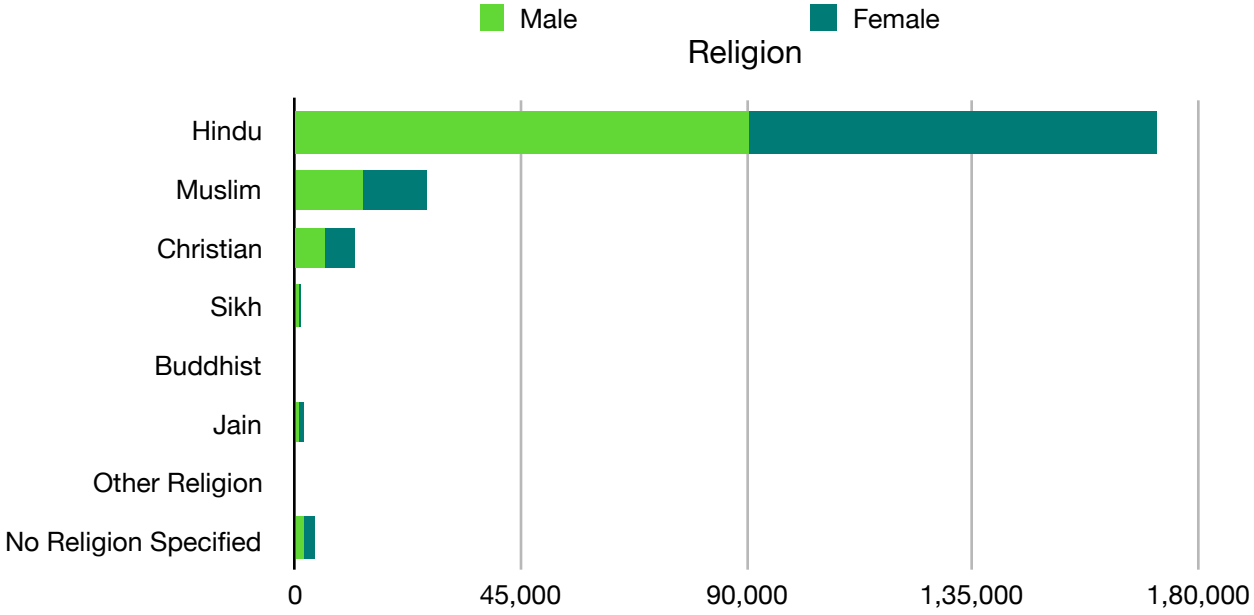
After collecting all the necessary data, the next step was to understand the data collected and draw conclusions from it. Graphs such as scatter plots, bar charts, line plots etc. were plotted initially to figure out some general trends in the data against time. This would also help to determine the factors suitable to include in the predictive model. With the help of these plots, various trends regarding the voter population, the winning party, the margin of votes between the winner and

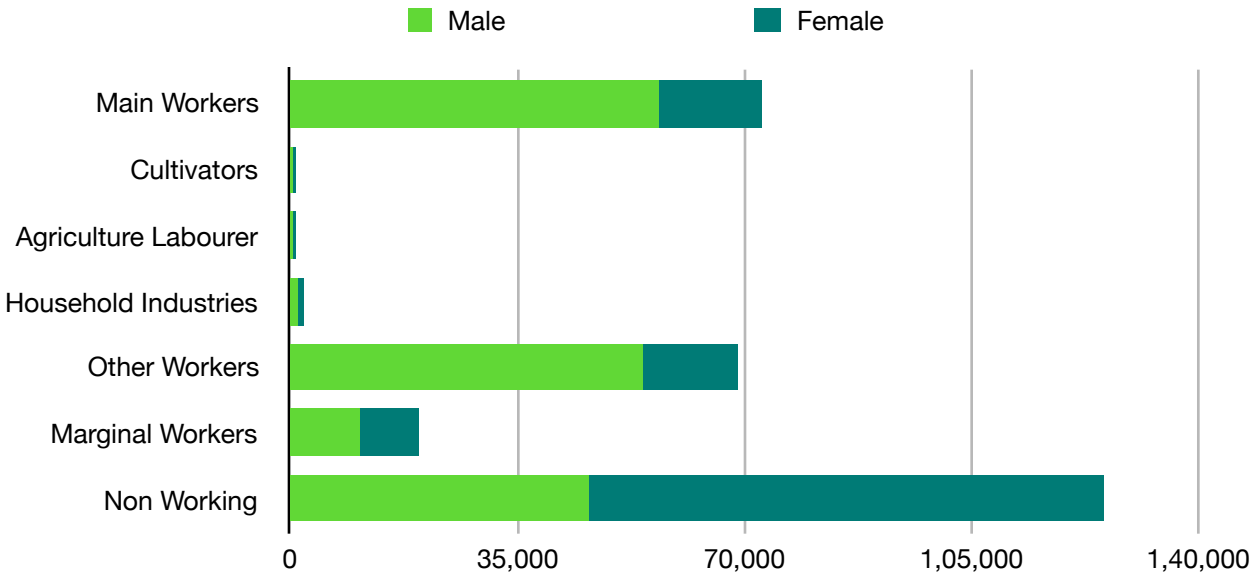
runner-up, and demographics of each constituency were analysed. Below are some visualisations of Secunderabad constituency.

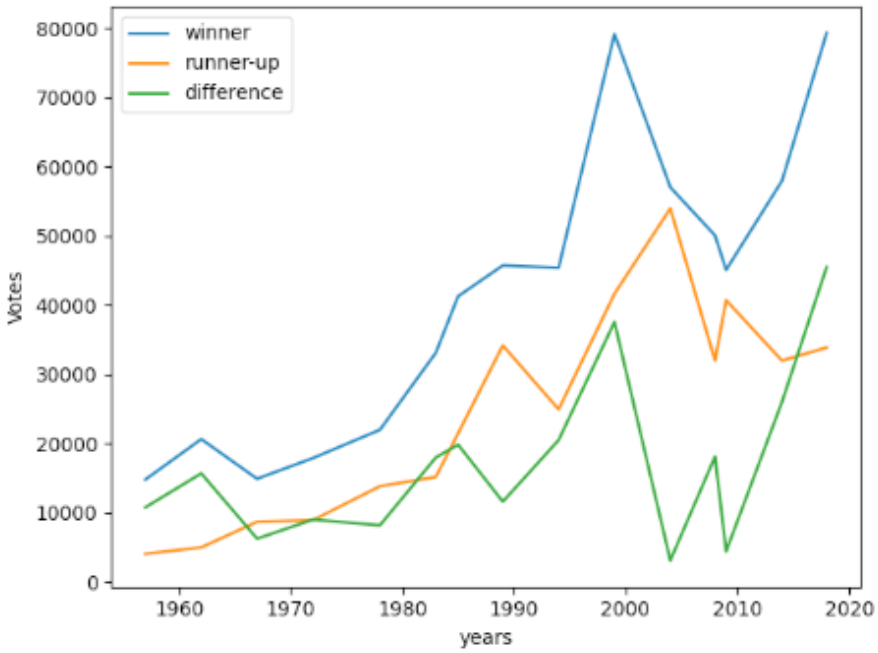





Male
Female
Scheduled Caste
Schedules Tribe
0
7,500
15,000
22,500
30,000
Number of people


Male
Female
Religion
Hindu
Muslim
Christian
Sikh
Buddhist
Jain
Other Religion
No Religion Specified
0
45,000
90,000
1,35,000
1,80,000


Male
Female
Main Workers
Cultivators
Agriculture Labourer
Household Industries
Other Workers
Marginal Workers
Non Working
0
35,000
70,000
1,05,000
1,40,000



Model Building

The predictive model which was built was based on Linear regression. For various parameters like demographics, candidate history, previous election results, the model was built by assigning coefficients obtained.

Demographic data was included in the Regression Model, but it gave muddy result with very low accuracy, hence the demographics data such as literacy rate, religious population distribution etc was dropped from the model (we ran this for Sivaganga - Lok Sabha Election, here data was taken from Census 2011 and 2001 and the points were extrapolated using a simple linear population growth formula). Another reason for lack of inclusion of the demographic data was that Shrug had the data for district(which includes parts of multiple LS constituencies) and it was difficult to relate this data of district level to smaller Assembly Constituencies. Hence, in the final model ran on candidate data and previous election results for the assembly constituencies.

For the implementation of the model, Python libraries such as Sklearn, Matplotlib, Seaborn, Pandas and Numpy were used. First, the data file was imported into a dataframe and the required factors (columns) were filtered out. The preditive

model, determines the percentage vote share of the winner for an election. The predictive factors were enop, n_cand, party, turnout_percentage, candidate_type, deposit_lost and vote_share_percentage was predicted. The data was split into training and testing data. Through this, we determined the coefficients of each predictive factor and analysed which factors were more significant than others. The accuracy obtained from this model for Secunderabad was 87 percent.

```
0.8753250540371812
                    Coefficient
tr_ac_name            -0.157248
n_cand                -0.066991
turnout_percentage     0.103280
enop                  -0.321406
party                  0.020264
deposit_lost         -31.653555
```

Future Prospects of the model:

By training the model with more urban constituencies, the accuracy of the model can be improved. With a higher accuracy, the model can act as an aid political parties in effectively choosing between prospective candidates depending on the input parameters of the candidate. Also, inclusion of demographics into the test data can be worked upon.

Key learnings from the project

- Election process in India

  Election is a very complex process for every democratic country. We have statistics that the process depends on many factors classified under quantifiable and non quantifiable. Few examples include literacy rate, party criminal cases against contesting candidates under quantifiable and

other back activities like bribes or booth capturing comes under non quantifiable. Each of us have tried to learn the history of elections in the constituencies provided to us.

- Factors affecting the win of a candidate

  There are numerous factors that affect the success of a candidate. Demographics of the constituency, age, gender, category, religion, party of the candidate, opinion polls, the statements given by the candidate, election expenditure, policies brought up by the candidate in the past and his popularity amongst the public.

- Data collection

  Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

- Data cleaning techniques

  Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. Data cleaning is not simply about erasing information to make space for new data, but rather finding a way to maximize a data set’s accuracy without necessarily deleting information.

- Data Visualization techniques

  Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

- Machine Learning Models and Algorithms

  Gained valuable insight and knowledge into the various algorithms used in ML models such as Univariate and Multivariate Linear Regression, Decision Trees for classification and regression, Random forests, other popular classification algorithms such as K-Nearest Neighbors, Naive Bayes, Adaboost and Linear Support Vectors Classifier.

- Working on a project as a team

  Collaborative efforts yield significantly better outputs as compared to individual efforts, collaboration helps individuals develop effective communication skills, and build good professional relationships.

- Approach to Data Science project

  The very first step in any project is defining the problem statement so is it for a Data Science project. It is followed by collecting relevant data from various sources or by conducting surveys. Data collection in the most important and tiring part of this whole journey. Once you have a good dataset, the next few steps are quite interesting - exploratory data analysis, data cleaning, outliers removal and normalization. Once all this is done, your dataset is ready for modelling. In this particular project we have to predict the election result outcomes and hence a regression algorithm is preferred over classifiers. Lastly, the model is evaluated and optimized for best results.

What courses in BITS have helped you in this internship

1. Probability & Statistics
2. Applied Stochastic Process
3. Statistical inference & applications
4. Machine Learning

Learnings:

- Working on a project as a team and moving forward considering everyone's opinions
- Election process in India
- Approach to a Data Science project
- Machine Learning Models and Algorithms mainly Multivariate Linear Regression and Time series Regression
- Data Visualization techniques using bar charts, scatter plots, line plots etc
- Data cleaning techniques/wrangling

References (for the project)

- Information on candidates https://www.myneta.info/
- Census data https://censusindia.gov.in/2011census
- Data on affidavit submitted by candidates https://affidavit.eci.gov.in/
- SHRUG data repository http://www.devdatalab.org/shrug
- Census data https://censusindia.gov.in/2011census
- Electoral Data https://eci.gov.in/ https://www.electionsinindia.com/
- Python Scikit Tutorial https://www.dataquest.io/blog/sci-kit-learn-tutorial/
- R regression analysis tutorial http://r-statistics.co/Linear-Regression.html