Analyzing/Forecasting Local Elections in Turkey

Mert Emin Kalender, 20702164

December 17, 2012
Outline

• Introduction
  • Project Understanding: Local Elections in Turkey

• Data Understanding
  • 2009 Local Election Results

• Data Preparation

• Modeling

• Evaluation

• Deployment / Demo

• Conclusion

• References
Introduction

- Analysis of people’s political choices
  - popular research area in political science
  - reasoning behind world views of individuals
  - voting, religious, consumer behavior etc.

- Turkish voters’ political choices
  - effect of personal values, political campaign etc.
  - after 2003, more analysis of election results
  - mostly about general elections

- This project focuses on local elections
Elections to choose local officials
- including metropoles, cities, various sized towns
  - 16 metropole, 65 city, 957 town, 1,974 small-sized town

Hold every 5 years
- recently some statistics of previous elections are published

Last one: March 29, 2009
- 3,000 officials selected

Next one: March 30, 2014
Introduction

Project Understanding

• Domain: Politics

• Benefit: An detailed analysis of election results
  • for each city based on following candidate attributes:
    age, educational background, marital status and political party

• Solution: A system predicting chance of winning
  • for a candidate given age, educational, marital, political party information

• Regression problem
  • value of interest is numerical
Introduction
Project Understanding
Turkish Statistical Institute (TUIK) published the data

A data set for each city

- with two tables: candidate statistics and winner statistics
- consisting of candidates for all local administrations of the city

Data is provided in different formats

- HTML, pdf and xls
- HTML is used and parsed with scripts

Example data set (for Ankara)

- age, educational background, marital status, political party
Data Understanding
2009 Election Results

- Syntactically and semantically **accurate** data
- No missing values
- No complex data types
- No need for dimensionality reduction
- No need for feature selection
- **Clean** and **simple** data sets
  - Is this enough?
Data Preparation

Step 1

- Given all candidate and winner statistics combined
- In total, 4 tables* are created
  - age, educational background, marital status, political party

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>City</th>
<th>Voter (%)</th>
<th>Candidate</th>
<th>Winner (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25-29</td>
<td>F</td>
<td>C24</td>
<td>6.033</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>30-34</td>
<td>M</td>
<td>C24</td>
<td>5.634</td>
<td>10.87</td>
<td>0.0</td>
</tr>
<tr>
<td>30-34</td>
<td>F</td>
<td>C24</td>
<td>5.145</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>35-39</td>
<td>M</td>
<td>C24</td>
<td>5.398</td>
<td>9.42</td>
<td>3.448</td>
</tr>
<tr>
<td>35-39</td>
<td>F</td>
<td>C24</td>
<td>5.265</td>
<td>0.725</td>
<td>0.0</td>
</tr>
<tr>
<td>40-44</td>
<td>M</td>
<td>C24</td>
<td>4.634</td>
<td>13.768</td>
<td>17.241</td>
</tr>
<tr>
<td>40-44</td>
<td>F</td>
<td>C24</td>
<td>4.483</td>
<td>0.725</td>
<td>0.0</td>
</tr>
<tr>
<td>45-49</td>
<td>M</td>
<td>C24</td>
<td>4.089</td>
<td>18.116</td>
<td>31.034</td>
</tr>
<tr>
<td>45-49</td>
<td>F</td>
<td>C24</td>
<td>4.127</td>
<td>0.725</td>
<td>0.0</td>
</tr>
<tr>
<td>50-54</td>
<td>M</td>
<td>C24</td>
<td>3.616</td>
<td>16.667</td>
<td>20.69</td>
</tr>
<tr>
<td>50-54</td>
<td>F</td>
<td>C24</td>
<td>3.561</td>
<td>0.725</td>
<td>0.0</td>
</tr>
<tr>
<td>55-59</td>
<td>M</td>
<td>C24</td>
<td>3.01</td>
<td>13.768</td>
<td>20.69</td>
</tr>
<tr>
<td>55-59</td>
<td>F</td>
<td>C24</td>
<td>2.936</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>60-64</td>
<td>M</td>
<td>C24</td>
<td>2.362</td>
<td>4.348</td>
<td>3.448</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Party</th>
<th>Sex</th>
<th>City</th>
<th>LocalR (%)</th>
<th>Candidate</th>
<th>Winner (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKP</td>
<td>M</td>
<td>C11</td>
<td>33.558</td>
<td>17.857</td>
<td>73.333</td>
</tr>
<tr>
<td>AKP</td>
<td>F</td>
<td>C11</td>
<td>33.558</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>DP</td>
<td>M</td>
<td>C46</td>
<td>4.139</td>
<td>5.97</td>
<td>8.065</td>
</tr>
<tr>
<td>DP</td>
<td>F</td>
<td>C46</td>
<td>4.139</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>CHP</td>
<td>M</td>
<td>C46</td>
<td>6.846</td>
<td>8.955</td>
<td>11.29</td>
</tr>
<tr>
<td>CHP</td>
<td>F</td>
<td>C46</td>
<td>6.846</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>DSP</td>
<td>M</td>
<td>C46</td>
<td>1.574</td>
<td>5.672</td>
<td>4.839</td>
</tr>
<tr>
<td>DSP</td>
<td>F</td>
<td>C46</td>
<td>1.574</td>
<td>0.299</td>
<td>0.0</td>
</tr>
<tr>
<td>SP</td>
<td>M</td>
<td>C46</td>
<td>3.644</td>
<td>8.955</td>
<td>3.226</td>
</tr>
<tr>
<td>SP</td>
<td>F</td>
<td>C46</td>
<td>3.644</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>MHP</td>
<td>M</td>
<td>C46</td>
<td>24.851</td>
<td>17.612</td>
<td>11.29</td>
</tr>
<tr>
<td>MHP</td>
<td>F</td>
<td>C46</td>
<td>24.851</td>
<td>0.299</td>
<td>0.0</td>
</tr>
<tr>
<td>BGZ</td>
<td>M</td>
<td>C46</td>
<td>0.062</td>
<td>0.896</td>
<td>0.0</td>
</tr>
<tr>
<td>BGZ</td>
<td>F</td>
<td>C46</td>
<td>0.062</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

* Data tables in reality are bigger than the listed ones. They are minimized for the sake of simplicity.
Various **regression** models applied [rm]

- **linear regression result for age data set** of Ankara
  
  \[
  0.186 \times \text{Age} + 1.553 \times \text{Sex} + 1.246 \times \text{Candidate} - 2.824
  \]

- predicts a win value between 4%-8% (with squared error > 50%)

- **linear regression result for educational background data set** of Ankara
  
  \[
  -3.016 \times \text{Education} + 2.950 \times \text{Sex} + 1.172 \times \text{Candidate} + 0.896
  \]

- predicts a win value between 12%-19% (with squared error > 55%)

- **similar results for other data sets** as well
Evaluation

Step 1

• **Biased models** obtained
  • Not enough data
  • Polynomial city information is a problem
    • city attribute expressed as `c[#plateNumber]`
  • Voter/LocalR features not really acceptable
    • values for the main city, (but each local place has its own)
    • removed with the expert consultation

• **No good results!**
Data Preparation

Step 2

- Data for each city separated from the data sets
- Data generation performed
- Each city with 4 tables* for
  - age, educational background, marital status, political party

* Data tables in reality are bigger than the listed ones. They are minimized for the sake of simplicity.
• **Regression** models (re)performed [rm]

  • Categorical values treated as ordinal

  • linear/polynomial regression result for age data set of Ankara
    
    \[-0.140 \times \text{Age} + 0.140 - 0.000 \times \text{Age}^2.000 - 0.376 \times \text{Sex}^2.000 + 0.168\]

  • linear/polynomial regression result for political party data set of Ankara
    
    \[0.013 \times \text{Party} - 0.137 \times \text{Sex} + 0.033 0.013 \times \text{Party}^1.000 - 0.134 \times \text{Sex}^4.000 + 0.033\]

  • logistic regression result for marital status data set of Ankara
    
    Bias (offset): -189.755
    \[w[\text{Marital}] = 240.427 \]
    \[w[\text{Sex}] = 222.459\]

  • logistic regression result for political party data set of Ankara
    
    Bias (offset): -14.888
    \[w[\text{Party}] = 420.013 \]
    \[w[\text{Sex}] = 37.619\]
• **ID3 Tree** performed on data sets [rm]

  • result for political party data set of Ankara (accuracy: ~90%, precision/recall: ~60%)

```
Party = AKP
| Sex = F: 0 {0=1, 1=0}
| Sex = M: 1 {0=18, 1=26}
Party = CHP
| Sex = F: 0 {0=1, 1=0}
| Sex = M: 0 {0=24, 1=11}
Party = DP
| Sex = F: 0 {0=6, 1=0}
| Sex = M: 0 {0=14, 1=1}
Party = MHP
| Sex = M: 0 {0=35, 1=7}
```

```
Party = ANP: 0 {0=7, 1=0}
Party = BBP: 0 {0=12, 1=0}
Party = BGZ: 0 {0=13, 1=0}
Party = BTP: 0 {0=45, 1=0}
Party = DSP: 0 {0=32, 1=0}
Party = DTP: 0 {0=7, 1=0}
Party = EMP: 0 {0=2, 1=0}
Party = HYP: 0 {0=2, 1=0}
Party = LDP: 0 {0=21, 1=0}
Party = MP: 0 {0=9, 1=0}
Party = SP: 0 {0=45, 1=0}
```
Modeling
Step 2

- **Naive Bayes** performed on data sets [rm]
  - result for political party data set of Ankara (accuracy: ~90%, precision/recall: ~60%)
Modeling

Step 2

• **RIMARC** applied on data sets [ri]

  • result for **political party data set** of Ankara

    Party (weight: 0.8378)
    If Party="AKP" Then Risk=0.5777778 (45)
    If Party="CHP" Then Risk=0.30555555 (36)
    If Party="MHP" Then Risk=0.16666667 (42)
    If Party="DP" Then Risk=0.04761905 (21)
    If Party="BGZ" Then Risk=0.0 (13)
    If Party="EMP" Then Risk=0.0 (2)
    If Party="TKP" Then Risk=0.0 (11)
    If Party="LDP" Then Risk=0.0 (21)

    If Party="HYP" Then Risk=0.0 (2)
    If Party="SP" Then Risk=0.0 (45)
    If Party="BTP" Then Risk=0.0 (45)
    If Party="DTP" Then Risk=0.0 (7)
    If Party="MP" Then Risk=0.0 (9)
    If Party="ANP" Then Risk=0.0 (7)
    If Party="BBP" Then Risk=0.0 (12)
    If Party="DSP" Then Risk=0.0 (32)

  • result for **age data set** of Ankara

    Age (weight: 0.3236)
    If Age="40-44" Then Risk=0.2037037 (54)
    If Age="45-49" Then Risk=0.17142858 (70)
    If Age="55-59" Then Risk=0.16666667 (42)
    If Age="50-54" Then Risk=0.14925373 (67)
    If Age="65-69" Then Risk=0.125 (8)
    If Age="35-39" Then Risk=0.10256410 (39)

    If Age="25-29" Then Risk=0.0 (22)
    If Age="30-34" Then Risk=0.0 (27)
    If Age="60-64" Then Risk=0.0 (16)
    If Age="70-74" Then Risk=0.0 (3)
    If Age="75+" Then Risk=0.0 (2)
• **Regression** still does not provide good results
  - still biased
  - *categorical* types changed into *ordinal* ones
    - some of them makes sense (age, educational background)
    - others do not (marital status, political party)

• **ID3 tree** and **Naive Bayes** provides some idea
  - about which values are significant and improve winning chance

• **RIMARC** results may be useful too
  - early to say something at this phase
**Data Preparation**

**Step 3**

- Dummy coding applied for all data sets

<table>
<thead>
<tr>
<th>ANP</th>
<th>MHP</th>
<th>LDP</th>
<th>DSP</th>
<th>DTP</th>
<th>IP</th>
<th>CHP</th>
<th>HP</th>
<th>BBP</th>
<th>BDP</th>
<th>MP</th>
<th>ODP</th>
<th>TKP</th>
<th>DP</th>
<th>SP</th>
<th>AKP</th>
<th>BTP</th>
<th>EMP</th>
<th>HYP</th>
<th>BGZ</th>
<th>Sex</th>
<th>Win</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

* Data table in reality is bigger than the listed one. It is minimized for the sake of simplicity.*
Modeling
Step 3

- **Neural Network** applied [rm]
  - with ten-fold cross validation and stratified sampling

- Four models are generated for each city
  - result for **age/educational background data sets** of Ankara
Evaluation

Step 3

• **Neural Network** provided good set of models
  
  • able to derive some meaning from the data
  
  • two output nodes: one for class 0 and one for class 1

  • enables us to say something based on output node values

• **Too few women** candidate/winner
  
  • 554 among 15,569 candidates

  • 26 among 2,931 winners

• **Another** Neural Network modeling without sex feature
Requirement of a way to combine four models

- How to combine four of them together and come with a result?
- **RIMARC** results may be helpful
  - performed again and **weights generated** by the algorithm saved

A web application created

- **Neural Network** models into ruby classes
- **RIMARC** results into feature weighting suggestions
- **statistics** from 2009 into various data graphs
  - [d3, jq, sn, tw] used

Link to the demo application on [hr]
Conclusion

- From the data sets of TUIK
- To highly capable a web application on local elections
- Clean and simple data is not enough
- Prepare the data in a such way that enables good analysis
- Data is (re)organized and (re)generated several times
- At the end in total (81 * 4) * 2 = 648 NN models created
- 81 different weight profile generated via RIMARC
- A good analysis of local elections is now possible
References

• [d3]: d3js (Data driven documents)
• [hr]: heroku
• [jq]: jQuery
• [rm]: Rapidminer
• [ri]: RIMARC
• [sn]: sinatra
• [tw]: twitter bootstrap
İbrahim Melih Gökçek is a Turkish politician who has been the mayor of Ankara since 1994. He won local elections in 1994, 1999, 2004, and 2009 with three different political parties. (image: http://forum.paticik.com/thumbnails/333/3db/06c/910/8d5/199/930/658/93d/5d1/44_450xNULL.jpg)