X-Education
- Potential Lead Identification

Submitted by – Soumik Banerjee

Problem Statement

- Given a potential leads profile data, identify the possibility of conversion of the same based on a score based profiling. The expectation is to identify potential leads and the expected accuracy is 80%

- **Available Information**
  - Candidate's demographic data
  - Website activities and timeline
  - Lead Quality and associated scores
  - Lead Source and Ad source data

- **Process**
  - Create a regression model based on the data
  - Output from the model should be a score [0,100] associated to each candidate
  - Threshold scores are determined to identify 'hot leads' based on various business requirements

Data Model – Feature Elimination

- Of the Available variables the following features were eliminated. The reasoning is as provided below:

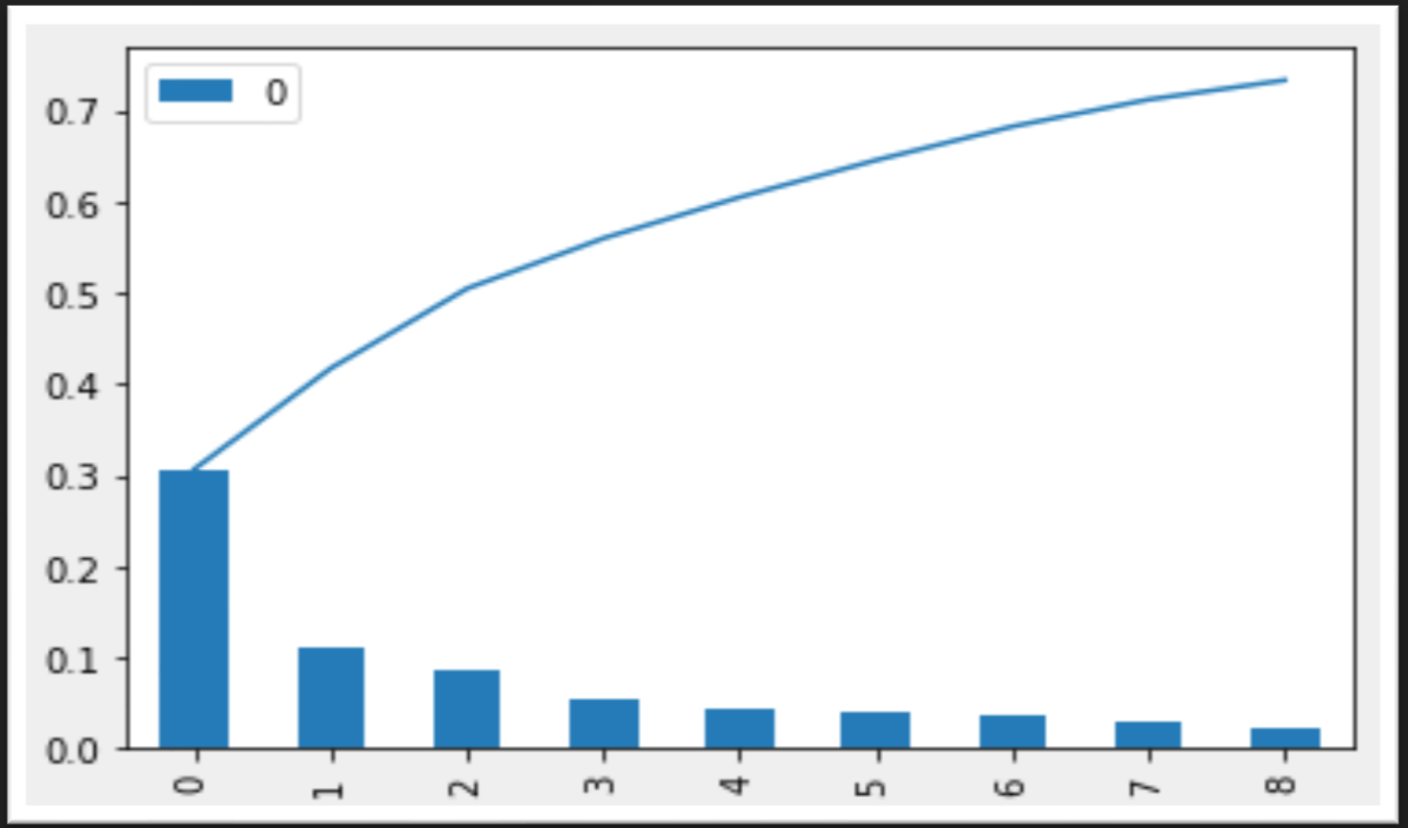
Feature	Reason of Exclusion
How did you hear about X Education, Lead Profile, Asymmetries Profile Score, Asymmetries Activity Score, Asymmetries Profile Index, Asymmetries Activity Index	Very High Null Percentage
City, What matters most to you in choosing a course, Do not call, Prospect ID , Magazine , Receive More Updates About Our Courses , Update me on Supply Chain Content , Get updates on DM Content , I agree to pay the amount through cheque	Lack of Data Diversity
Lead Origin, Last notable activity	Dependent variables - have no independent data quality

Data Model – Principal Components

- Dimensionality Reduction technique was used to ensure that all features are included in the model instead of Feature Selection (RFE).
- PCA analysis was performed on the data and variance of ~95% was achieved using 34 components.
  - The resultant model had 96% AUC score.
- Upon further experimentation, with only 9 components, which together gave 75% data variance, the model had 92% AUC score.
- Our final model had 9 Principal Components.

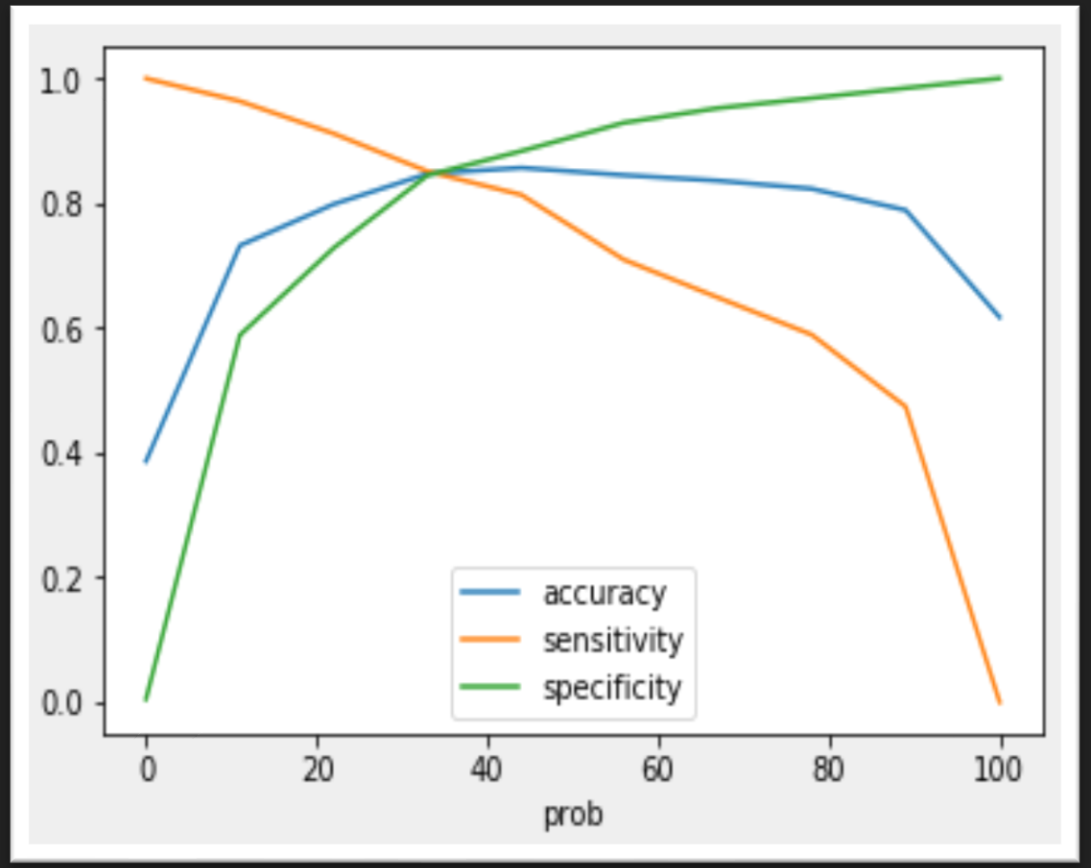
Data Model – Principal Components
(Scree Plot)

- **Scree plot** shown in the picture shows that the first components accounts for ~30% of overall data variance
- The other components, individually account for less than 10% of data variance



Data Model

- The Logistic Regression model fitted into the available data set gives the probability of conversion of each candidate.
- The score is determined by multiplying the generated probability by 100.
- Based on the accuracy vs Sensitivity vs Specificity plot, the threshold score is selected as 37.
- Threshold of 37 gives
  - 84.7% accuracy,
  - 83.3% sensitivity and
  - 85.6% specificity



Business Scenarios and Associated Scores

- For Aggressive marketing, during intern hiring session –
  - Increase sensitivity i.e., increase the chances identifying the positive cases more while compromising on accuracy. Based on the model, the following is suggested:
  - **Reduce the threshold score to 13**. It gives 75% accuracy but 95% sensitivity

- For Lean periods, when only high potential leads are to be contacted-
  - Increase Specificity, i.e., increase the chances identifying the negative cases more while compromising on accuracy. Based on the model, the following is suggested:
  - **increase threshold score to 66**. gives 95% specificity with 84% accuracy

Business Insight and Recommendations

- The model is based on dimensionality reduction principles and hence all constituent variables are influencing the model. Based on the Components selected in the model, the following variables are observed to have greater influence the component behavior. Hence **the following variables indicate higher lead conversion probability.**
  - Website visits (includes Page views/Total number of visits/Time)
  - Lead Quality
  - Last Activity

- The following categorical variables are identified to be influencing the selected components. These variables **are indicators where X-Education can focus to achieve better Lead Conversion:**
  - Lead Source
  - TAG
  - Occupation