EXPLORATORY DATA ANALYSIS

Brief description of the data set and a summary of its attributes

The data I took is about hiring new associates/ salespeople by managers in sales division. Since the company is making upfront investment in hiring and training, it tries to predict based on past data if an applicant will produce business after getting hired in future.

So, the data is made up of 14572 rows (each of a distinct applicant) and 22 features for prediction and one binary target variable of which training data has 9527 entries, and test has rest. So we have 14572 people who applied in the past and got hired and whether they sold a product within three months is reflected in business sourced target column. The features/ variables are:

- Identification: 'ID'
- Location of Offices: 'Office_PIN'
- Applicant Details: 'Application_Receipt_Date', 'Applicant_City_PIN', 'Applicant_Gender', 'Applicant_BirthDate', 'Applicant_Marital_Status', 'Applicant_Occupation', 'Applicant_Qualification'
- Manager details: 'Manager_DOJ', 'Manager_Joining_Designation', 'Manager_Current_Designation', ‘Manager_Grade', 'Manager_Status', 'Manager_Gender', 'Manager_DoB', 'Manager_Num_Application', 'Manager_Num_Coded', 'Manager_Business', 'Manager_Num_Products', 'Manager_Business2', 'Manager_Num_Products2'
- Target: 'Business_Sourced'[1]

Here, variable types are:

- Date: 'Application_Receipt_Date', 'Applicant_BirthDate', 'Manager_DOJ' and 'Manager_DoB'
- Numerical: 'Manager_Num_Application', 'Manager_Business', 'Manager_Num_Products', 'Manager_Business2', 'Manager_Num_Products2'
- Categorical: rest all

Initial plan for data exploration

1. Handling date variables
2. Handling non-numerical variables
3. Handling missing values
4. Handling outliers
5. Variable Transformation/ Feature engineering

Actions taken for data cleaning and feature engineering

1. Handling date variables
    a. Since we cannot handle date variables directly, we can get the information out of it by creating new columns.

[1] For the rest of the document, successful means Business sourced is 1, otherwise 0.

b. To see if month or year of application date has any effect in determining target, creating separate columns each for month and year instead of date directly

c. From applicant’s DoB, creating a variable for applicant’s age since I believe month and date wont have any meaningful impact on their success of work

d. Similarly using 'Manager_DOJ' and 'Manager_DoB' to calculate manager’s age and experience in the company

2. Handling non-numerical variables:

   a. PIN codes were used as geocodes to identify distance between applicant address and office location.

   b. Gender and Manager Status (whether probation or confirmed) were binary encoded.

   c. One Hot encoding (encoding with indicator variables) was performed for: 'Applicant_Marital_Status’, 'Applicant Occupation'

   d. Label encoding (Rank ordering) was performed for: 'Manager_Joining_Designation', 'Manager_Current_Designation', ‘Manager_Grade'. Grade was already ordinal; designation was of format “Level xx” was converted to ran based on xx.

   e. Target encoding was performed for: 'Applicant_Qualification'

3. Handling missing values

   a. I noticed something in the data wrt missing values, ie, for 683 observations there was no information on any column describing manager detils. So I figured these entries must mean, they had no manager, So created a new column for indicating the presence of manager.

   b. Features having missing values and their count in train apart from these were:

i.	Applicant_City_PIN	80
ii.	Applicant_Gender	53
iii.	Applicant_BirthDate	59
iv.	Applicant_Marital_Status	59
v.	Applicant_Occupation	1090
vi.	Applicant_Qualification	71

   c. Since everything except birthdate was a categorical variable and #missing is less than half a percent of overall data, observations with birthdate missing were dropped and others excluding Applicant_Occupation were imputed with KNN iteratively with Applicant_Gender having least missing imputed first. Due to large number of missing values, Applicant_Occupation missing was treated as a new category.

4. Handling outliers

   a. 'Manager_Num_Products' had unproportionally large count of 0 and very low counts of extremely high values. Hence, they were floored and capped at 10, 90 percentiles repectively and indicator for flooring and capping was added as a new feature

   b. 'Manager_Business' also had a couple of negative values and some extremely high values. Hence, they were capped and floored at 95, 5 percentiles respectively.

5. Variable Transformation

   a. 'Manager_Business' was heavily skewed towards left hence a log transformation was applied.

   b. 'Manager_Business2', 'Manager_Num_Products2' are essentially second line of business of a manager if exists. So, when it doesn’t the data replicates the first and replicates them which leads to extremely high correlation between them. So instead of these, their difference from 1st line of business, ie, 'Manager_Business', 'Manager_Num_Products' were calculated and used to create new columns and dropping these.

   c. A new column for growth of manager was calculated as a difference of 'Manager_Joining_Designation' & 'Manager_Current_Designation'

   d. Applicant_Qualification was re-grouped into Class X, Class XII, Graduate, Certified Associate, Masters & Others and target encoded

Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner

- Based on target event rates, Female applicants are lesser in number but more successful.
  - Number of male and female applicants are 11079, 3404 but proportion of applicants doing successful work is 33.2%, 38.5% resp.
- Managers are overwhelmingly male.
  - Number of male and female managers are 11321, 1744 but proportion of applicants doing successful work is 33.6%, 38.4% resp.
- Students tend to be more successful than other applicant occupations
  - Students success rate is 48% while others' range from 33% to 37%.
- Managers with higher Joining Designations tend to be more efficient in hiring more successful salespeople
  - Mangers with Joining Designations of Levels 4,5,6,7 have success rate between 50 to 62% while managers in lower designations have success rate between 22 and 36%
- Similarly, Managers with higher grade tend to be more efficient in hiring more successful salespeople
  - Mangers with grade above 6 have success rate between 50 to 57% while managers upto grade 6 have success rate between 12 and 42%

Formulating at least 3 hypotheses about this data

- Applicant Qualification has no effect in success of salespeople
- Number of products under a manager increases with managers business
- Number of applications to a manager decreases with that manager's business
- Students are more successful in becoming good sales agents than other applicant professions

Conducting a formal significance test for one of the hypotheses and discuss the results

Testing for if Students are more successful in becoming good sales agents than other applicant professions or not:

Null hypothesis: Students are as successful as other applicant professions

Vs

Alternate hypothesis: Students are more successful as other applicant professions

Let p1, p2 be proportion of successful applicants in student and non-student professions.

Then, H_0: p1 = p2 vs H_1: p1 > p2.

I tried two different tests:

1. Fisher's Exact test: Odds ratio = 0.55, p-val = 0.003
2. Pearson's chi-square test: Chi-sq stat = 4787.38, p-val = 0.0001

With both tests having p-value of less than 5%, we can conclude that the test is significant hence, students are more successful as other applicant professions.

Suggestions for next steps in analysing this data

After this data cleaning and feature engineering this data is ready for a classification model to be build on top and with successful model fit, test data can be used to predict the target aka Business Sourced

A paragraph that summarizes the quality of this data set and a request for additional data if needed

The collection of variables for this business problem was good per say but a lot of pre-processing work was needed to make this data fit for model building.

There was quite some number of entries where all manager details were empty. I don't know if that was just coincidence or different segment of people employed or just data entry error. No information was provided for that.

Data on which attempt in which the applicant passed and the score that they got in training during hiring would have been great in determining whether an applicant would generate business in future or not.