HR PREDICTING MODEL
AUGUR

FOUNDATIONAL PROJECT

CRISP-DM REPORT BY:

ARCHIT SRIVASTAVA - 12120052

DIPTI ASHOK BELURGIKAR - 12120099

JATIN SAHNAN - 12120098

KARTIKEY AJAY RAI - 12120014

SWATI YADAV - 12120084

BUSINESS AND DATA UNDERSTANDING

Business Problem:

Recruiters undergo complex procedures to hire candidates. Searching through various job profiles for the right candidate is a tedious job and goes to waste when the selected candidate rejects the offer at the end.

Business Objective:

The objective of this project is mainly to reduce the hassle of recruitment process by applying the concepts of CRISP-DM to the current system. We develop an Automatic Tracking System (ATS) that can map the resume of a candidate to the job description and then predict whether h/she would accept or reject the job offer.

Business Constraints:

Attributes influencing acceptance are different for each candidate. Each candidate has different factors such as expected salary, current salary, age, experience, etc. with which it becomes difficult to predict who would accept the offer.

DATA HANDLING AND MODEL CONSTRUCTION

To achieve our business objective, we have created a Streamlit application named Augur that enables the user to get the desired result in a seamless and efficient way

The entire application is derivation of three-pronged model:

1) Creation of Skillset list which serves as basic input for the application:

- We are scrapping different job description on Indeed.com and are using NER to extract skills for our skillset list which will be used in our model.
- First, BeautifulSoup in Python is used to extract the text from data analyst profiles from the indeed.com URL and save it to a variable named 'text'.
- We then create a function for the converting this text into JSON using NER. We pull out the text and entity from the 'text' above and assign them as key value pairs in a dictionary.
- We finally print out this dictionary as a data frame.
- This code is then applied for different profiles within the Data Analyst domain.

2) ATS model giving score to candidate on basis of skillset dictionary or JD uploaded as input

- Since our business objective is to minimize the cost of recruitment process, we aim to create a model for the HR Department.
- This model assigns weighted scores to each candidate in the HR dataset with respect to the skills possessed by him/her which goes as an input to the prediction model to predict whether a shortlisted candidate will accept or reject the offer.
- We mainly use pandas, sklearn and nltk libraries for our ATS model.
- Initially, the sample dataset is imported by reading it into a data frame. We see that the Resume is input as a column in this data frame. Other columns present are Experience, Offered Salary, Expected Salary, Current Salary, etc.
- Basic cleaning steps are done by dropping the null values and removing special characters from the resume column text.
- Next, all the text in resume is converted into lower case for better accuracy on the string operations.
- We then do a split by word for this column and assign it to a new data frame. We calculate the frequency for each word.
- This data frame is then filtered for the key words required as per job description, such as Hadoop, Big Data, Visualization, AWS, Power BI, etc.
- We fill all the NAs with 0s, and then calculate the Total Score by adding all the column values.
- This Total Score column is finally merged with the original data frame to show the Total Features score for each candidate.
- This data frame is now exported as a csv file called 'skills_resume.csv'.

3) Prediction model which predicts whether Candidate will accept or decline, based on the input variables defined in Streamlit app. We use Streamlit to provide an attractive and smooth, hassle-free experience for the recruiters.

- The Home Page is divided into three sections:

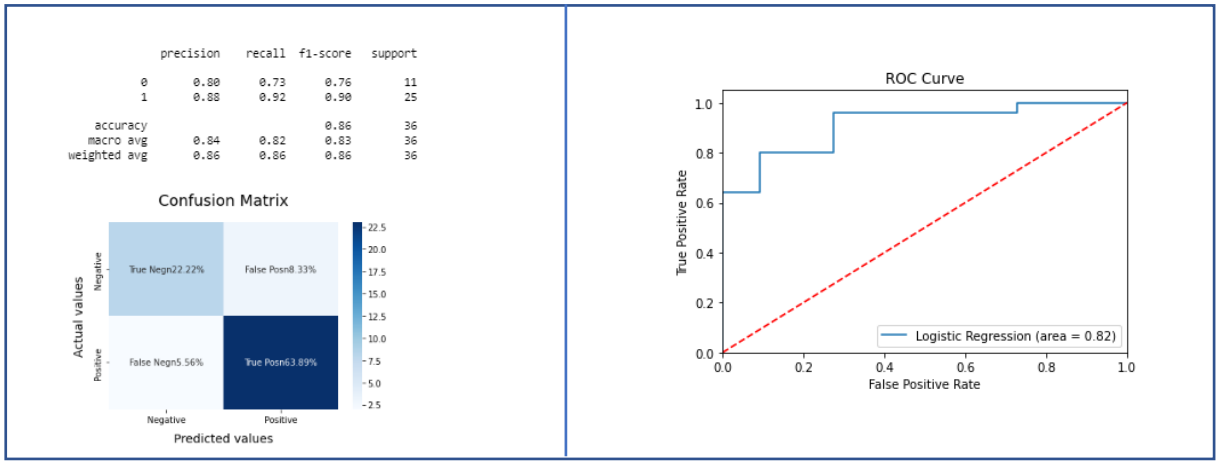
- Selecting or Uploading Job Descriptions – The recruiter is required to select or upload the job descriptions or skills needed in the candidate they want to choose.
- Other information required: They need to fill some more information regarding the salary of the candidate, location specific details and experience.
- Resume Upload: The recruiter then uploads the resume of the shortlisted candidate which they wish to know if he or she will accept or reject the offer.

- The next page simply provides a print of the Resume uploaded in the Home page such that the recruiter can review for any details.
- Using NLP techniques, we search whether the job description uploaded by the recruiter appears in the candidates resume and assign a frequency count against each word. We also calculate the Total score of this candidate.
- In the next page, we use the dataset received by the HR and search for the job description in this dataset to find how many candidates are applicable for the given JD.
- The last page shows whether the candidate has accepted or rejected the offer. To determine this our prediction model uses Logistic regression.

    - For training our prediction model, we take in the output data frame of the ATS model mentioned above. We only take specific columns i.e., Experience, Current Salary, Expected Salary, Offered Salary, Current Location, Offered Location, Offer Status and the Total Score (Named as Feature_Total in our data frame)
    - A quick check is done for the balance of data based on the Offer Status. We see that percentage of Accepted is 61.6% and percentage of candidates declines is 38.33%.
    - Next we create two new features as follows:
      1) Loc_Bin: Is the current location and job location same, yes = 1 & no = 0
      2) Offer_Bin: Offer Accepted = 1; Offer Declined = 0
    - We then fit our model using Logistic regression with y being our Offer_Bin column and x being the rest of the columns : 'Exp (Yrs)','Current Salary', 'Expected Salary', 'Offered Salary', 'Feature_Total', 'Loc_Bin'. After fitting, we see that the p-value is 6.7131e-14
    - We then split our data into 70:30 ratio, with 30% being the test data size and then apply logistic regression on the training data.

MODEL ACCURACY:

- The accuracy of our model is 0.86 which is a good number. We also plot a confusion matrix which showcases high true positive percentage of 63.89% and our logistic regression ROC curve area is 0.8



Using this model, we take the candidate's attributes and check the prediction on whether this candidate accepts or rejects the offer and print it on the screen.