PREDICTING THE RENEWAL PREDICTION OF THE MEMBER

HEALTHCARE INSURANCE POLICY RELATED

BY JAYANTHI ELUMALAI

AGENDA

- Business Problem statement
- Data Acquisition & Cleaning
- Exploratory Data Analysis
- Predictive Modeling
- Result & Conclusion
- Lesson learnt during project time
- Future Enhancements

BUSINESS PROBLEM STATEMENT

- Heath Insurance company always want to know the member renewal rate and how much to charge member according to the rules provided by the goverment. Members will renew the policy every year (Covering only the Individula policy whoes having the default pricing for family member till 5 and if the members of the family is more than 5 then Premium will get increased accordingly.

DATA ACQUISITION & CLEANING

- Data acquired from the Kaggle for insurance and other needed data where added manually to the dataset.

  1) Policy detail and member details joined together and retrieved the data

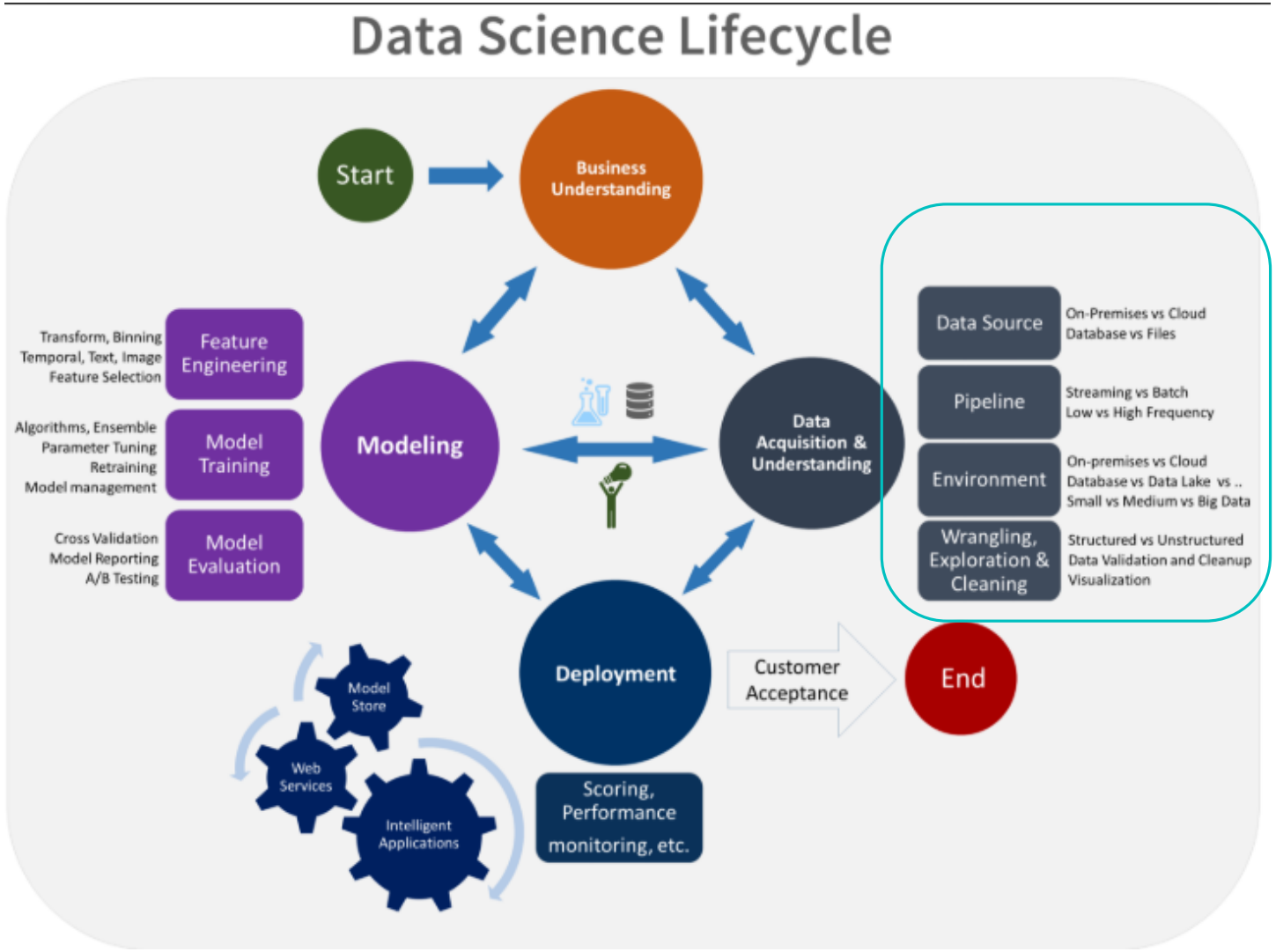
  2) Corporate\Small group policy where grouped together and removed from the data set, since this prediction if only for the individual policy.

  3) Member ID , DOB , SSN and other Secured information should be removed before transfer the data from the DB.

  4) Member Name and dependent name Policy Number is not needed for creating a Machine Learning model hence removed those.

  5) Plan, Region and county and some other fields where categorical so need to change those using Label encoding and One Hot encoder

- After fixing these problems, I checked for outliers in the data. I found there were some extreme outliers, mostly caused by some types of small sample size problem. Example : Member is very new and he is not having any renewal before.



EXPLORATORY DATA ANALYSIS

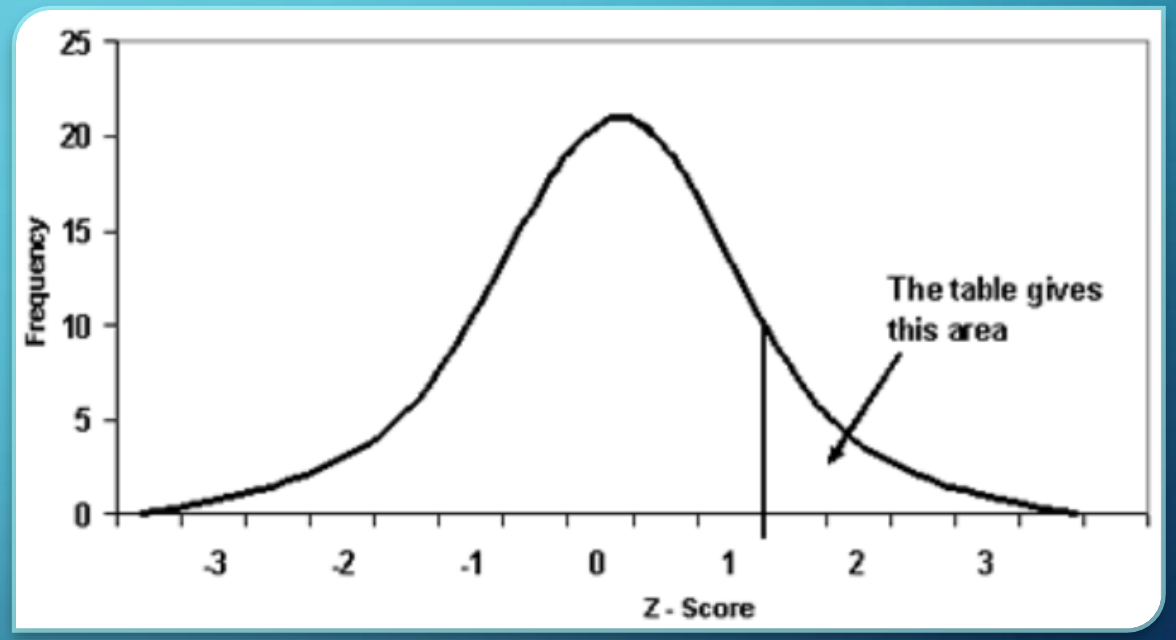
- I have used LabelEncoder and OneHotEncoder
- **LabelEncoder :** To convert the repeated Categorical\ text to Number for the fields like sex (Male, Female) , States, Plan (Gold, Platinum, Silver, Bronze)
- **OneHotEncoder :** To convert the single common numerical data to multiple columns with 1's and 0's , Dataset Shape - column count will get increased.
- We have to use the above EDA process to make the model process fast with numerical data..

PREDICTIVE MODELING

- In Dataset, there are more categorical variable which is impacting the target variable, Hence I have chosen the Multiple Linear Regression (MLR) model with **Backward** and **Forward** selection process to get the best categorical fields which will provide the accurate result.
- Initially I have set the SL value as 0.05 and the running the model each time By checking the p_value remove the highest p_value categorical field and check score value which is getting improved.. So followed the same steps to get the best score using the backward and forward selection processing.

RESULT & CONCLUSION

- Using the Linear Regression with multiple Variable's we were able to get the prediction with ~95% accuracy after 7th iteration, Since there is a possibility of the improvement by avoiding the seasonal purchase of policies, Hence concluding that we can enhance this in future for better prediction.



LESSON LEARNT DURING PROJECT TIME

- **Compatibility issue's:**
  - I was using only the IBM Watson studio Notebook for doing the exercise due to which i was facing some compatible.. Notebook has the default python kernels and its packages. I learnt how to use the Environment with different version of python which will solve the compatibility between the packages.
- **Asset Handling:**
  - I was using the existing dataset from the IBM cloud, But when it comes to project I created my own file and loaded to Data asset, But i find very difficult to include the file to the code.
- **Dataset Creation:**
  - Initially I was using the Insurance dataset from Kaggle site, later found I need some more fields need for my dataset which Insurance start date and end date, etc.,
  - For Which I used excel formulae to generate the date's using Choose() and Randbetween() methods
- **Different Algorithms:**
  - To choose the best algorithm for the project exercise, I have browsed more and came to know we can apply ARIMA model and some Deep learning model also for this problem

FUTURE ENHANCEMENTS

- In Dataset, I have the date field for Policy join date and Policy end date, Using which we can identify the different trends with ARIMA (Auto-regressive integrated moving average) like Christmas season Sale and other specific trend with historical data Using the Notebook file or PKL file integrated with the REST API and provide model prediction output to the end user applications

JAYANTHI ELUMALAI

For Coursera Final assignment (IBM DATA SCIENTIST CERTIFICATION)