

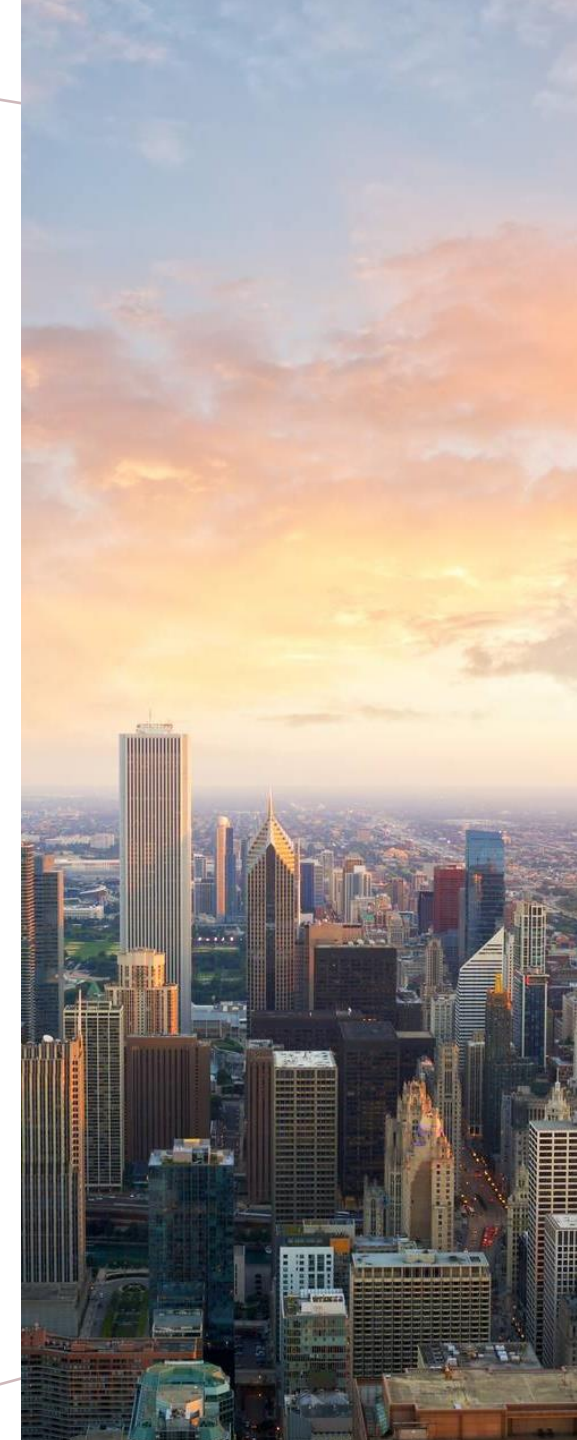
LEAD SCORE CASE STUDY

*BY ISAAC MARTIN, JAI GANESH G &
JAYDEEP GOKHALE*



Agenda

- Project Summary
- Approach
- Methodology
 - Data Understanding
 - Data Quality
 - Data Preparation
 - Model Development
- Outcome
- Recommendation



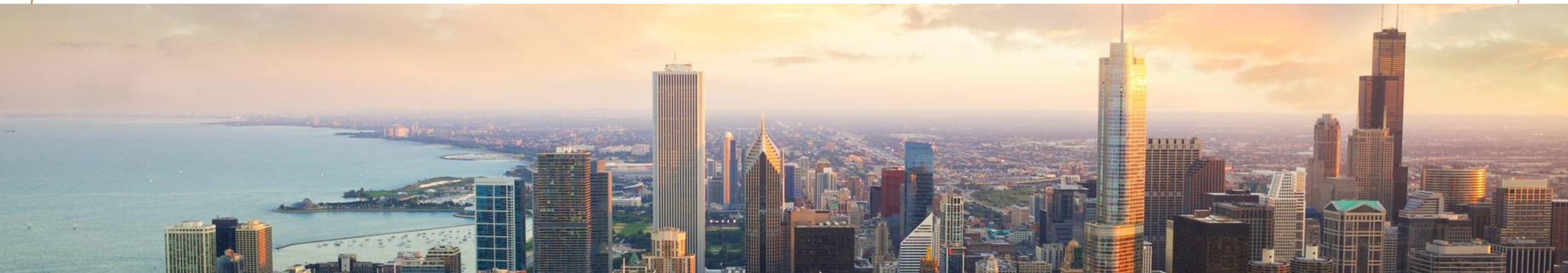
PROJECT SUMMARY

1. OUR SALES TEAM ATTRACTS PROSPECTS TO OUR COMPANY WEBSITE VIA MARKETING ACTIVITIES .
2. INTERESTED PROSPECTS FILL IN AN ENQUIRY FORM AFTER WHICH THEY QUALIFY AS A LEAD.
3. THE SALES TEAM CALLS UP ALL THE LEADS, HOWEVER CONVERSION IS CURRENTLY ONLY AT 30%. 70% OF THE SALES TEAM EFFORT IS UNPRODUCTIVE.
4. PROJECT MAX AIMS TO REDUCE THIS NON - PRODUCTIVITY AND INCREASE THE SALES TEAM CONVERSION RATIO.

QUALIFY LEADS BASED ON KEY FACTORS WHICH HELP PREDICT POSSIBILITY OF CONVERSION.

ASSIGN A CONVERSION PROBABILITY TO EACH LEAD.

SALES TEAM CAN NOW FOCUS ON HIGHER CONVERSION PROBABILITY LEADS TO IMPROVE PRODUCTIVITY AND CONVERSION.



APPROACH

- THERE IS A GOOD QUANTITY OF HISTORICAL DATA AVAILABLE
- THIS DATA WAS USED TO TRAIN A MACHINE LEARNING MODEL.
- CONVERTING BOOLEAN COLS TO NUMBERS TO MAP THE CORRELATION
- WE HAVE USED LOGISTICS REGRESSION SINCE OUR OBJECTIVE IS LEAD CLASSIFICATION
- THE MODEL GENERATES CONVERSION PROBABILITY SCORES FOR EACH LEAD
- THE SALE TEAM, DEPENDING ON THE PREVALENT BUSINESS CONDITIONS, MANPOWER AVAILABILITY AND SALES PRESSURE, CAN NOW DECIDE UPON A CUT OFF PROBABILITY AND ONLY CONTACT THE LEADS ABOVE THIS VALUE
- THE MODEL WILL BE RE-EVALUATED AFTER A PERIOD OF A YEAR OR EARLIER IN CASE THERE IS A SIGNIFICANT CHANGE IN ANY OF THE PARAMETERS






METHODOLOGY

DETAILS OF THE STEPS PERFORMED TOWARD MODEL BUILD

Available data was inspected to get a sense of :

- Variables available in the data 
- Type of data in each variable – Numeric, Categorical, Text etc.
- Distribution of values in each variable
- Basic statistical summary of numeric variables.

DATA QUALITY ISSUES & CORRECTIVE ACTIONS

1) Presence of default 'Select' value in most variables

Variables like 'How Did You Hear About X Education', 'Specialization', 'City' had anywhere between 25 – 60% of values as 'Select' .

2) Variables with similar information

Example, the variables 'Last Activity' and 'Last Notable Activity' contained almost same information .

3) Variable to profile the lead – created by salesperson after interaction with the lead

There were 7 such variables like 'Tags', 'Lead Quality' and 'Lead Profile' which were removed from the dataset as they would not be available for raw data.

4) Missing values

Variables – 'How Did You Hear About X Education', 'City' and 'Country' were 3 variables that were dropped on account of presence of almost 30% missing values.

DATA QUALITY ISSUES & CORRECTIVE ACTIONS

5) Categorical variables with large number of levels and most having very low population

6 Variables like 'Lead Source', 'Total Visits' and 'Page Views Per Visit' had upwards of 10 levels in each with most of the trailing levels having population under 3. All such bottom levels were combined in line with the overall distribution of values in each.

6) Outliers in numeric variables

Extreme outliers were found to be present in the field 'Total Time Spent on Website' however, considering that there could be a correlation between a serious lead who spends higher time on website and the probability of conversion we decide to retain the outliers

DATA PREPARATION

1) Test for Imbalance

No serious imbalance was observed as the split between converted and not converted at 37/63.

2) Creation of Dummy Variable

Dummy variables were created for all the categorical variables.

3) Correlation Analysis and Elimination of highly correlated variables

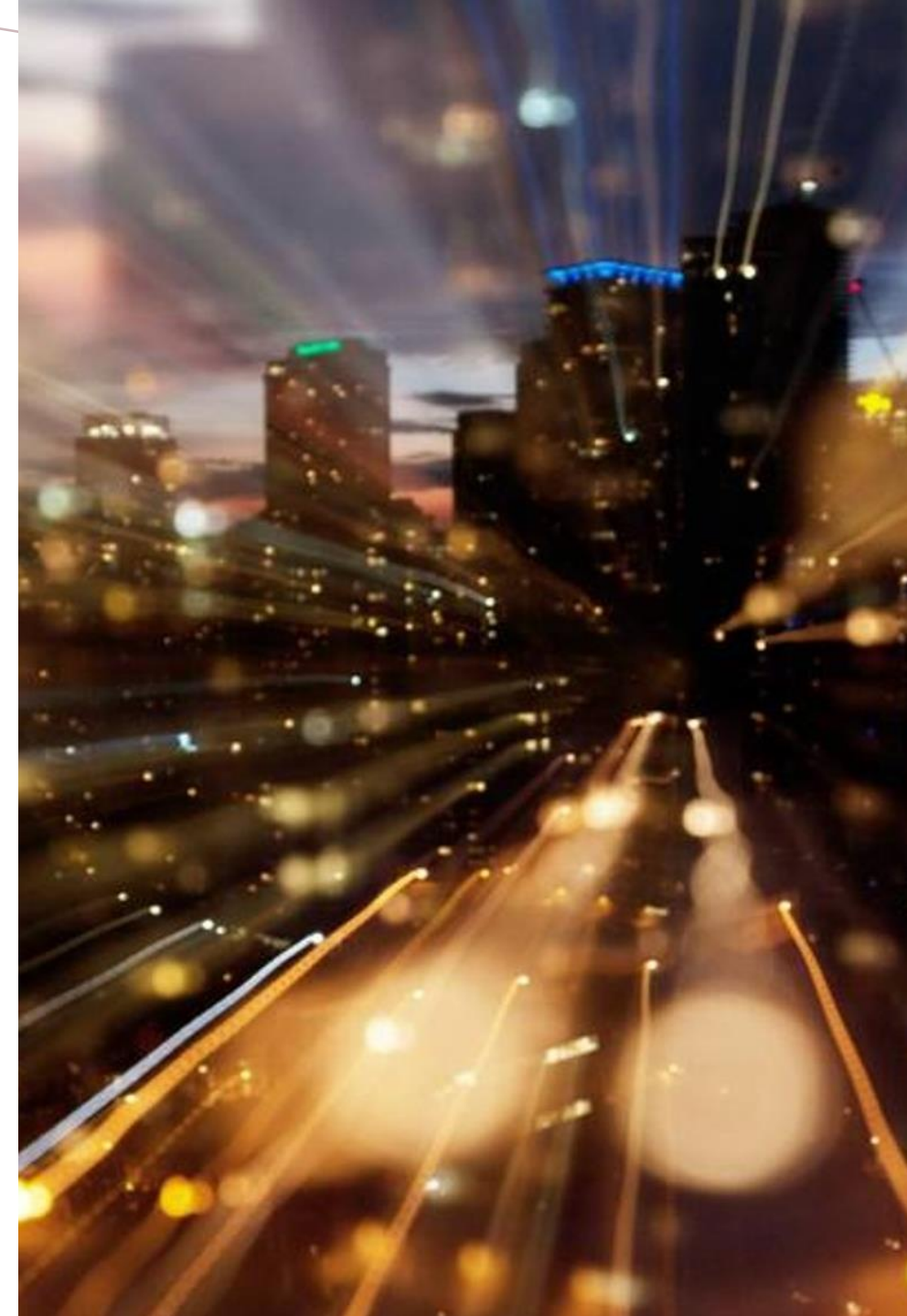
The correlation map was studied, and a decision was taken to eliminate the top 5 variables with correlation above 80% only. The rest would be taken care of by RFE.

4) Split data into train and test datasets (70/30 split)

5) Scaling of variables (Standard Scaler)

Scalar was fit on training dataset and used to transform both training and test datasets.

6) Converting Boolean cols to numbers to map the correlation



MODEL DEVELOPMENT

- We used RFE to reduce the number of variables from 44 to 25
- Starting with these 25 variables, successive models were constructed and the 'p' and 'VIF' Values were tested at each step to recursively eliminate nonsignificant variables.
- After 12 iterations, we got the final model with 13 variables and all parameters within prescribed limits
- ROC curve was plotted to check the AUC
- The model was then run on the Test dataset and verified that there was no significant variation between the performance parameters
- Values for Accuracy, Sensitivity and Specificity were generated at various probability level between 0 and 1 and the resulting curves plotted to obtain the ideal cut off probability of 0.3.

OUTCOME –MODEL PERFORMANCE

- We have been able to develop a high-performance model in line with best industry standards.
- The 'P' values for all variables were well below 0.02
- The 'VIF' Values for all variables were well below 3
- The performance parameters over both Train and Test sets were comparable, implying there was no overfitting:
 - Accuracy → train : 80 % , test : 79%
 - Sensitivity → train : 73% , test : 75%
 - Specificity → train : 84% , test : 81%
 - False +ve rate → train : 16% , test : 18%
 - Precision → train : 74% , test : 70%

OUTCOME – FINAL LIST OF VARIABLES

	coef	std err	z	P> z	[0.025	0.975]
const	-2.0018	0.087	-22.901	0.000	-2.173	-1.830
Total Time Spent on Website	1.3410	0.064	20.850	0.000	1.215	1.467
Lead Origin_Lead Add Form	2.7154	0.216	12.579	0.000	2.292	3.138
Lead Source_Welingak Website	2.1291	0.758	2.809	0.005	0.643	3.615
Do Not Email_Yes	-1.7727	0.178	-9.931	0.000	-2.122	-1.423
What is your current occupation_Other	2.5684	0.524	4.901	0.000	1.541	3.595
What is your current occupation_Student	1.2017	0.219	5.483	0.000	0.772	1.631
What is your current occupation_Unemployed	1.1180	0.085	13.212	0.000	0.952	1.284
What is your current occupation_Working Professional	3.6074	0.193	18.718	0.000	3.230	3.985
Last Notable Activity_Modified	-0.6212	0.081	-7.632	0.000	-0.781	-0.462
Last Notable Activity_Olark Chat Conversation	-0.8458	0.306	-2.766	0.006	-1.445	-0.247
Last Notable Activity_Other	1.8101	0.304	5.952	0.000	1.214	2.406
Last Notable Activity_SMS Sent	1.3384	0.083	16.120	0.000	1.176	1.501
TotalVisits_2.0	-0.3760	0.086	-4.394	0.000	-0.544	-0.208
Page Views Per Visit_7.0	-0.7305	0.277	-2.635	0.008	-1.274	-0.187

RECOMMENDATIONS

- **Calculate Conversion Probability Scores:** Utilize the model to determine the conversion probability score for each new lead. We will develop a user-friendly interface to facilitate this process for the sales team.
- **Understand the Minimum Cutoff:** The model's minimum cutoff score is 0.365. Leads with scores below this threshold are unlikely to convert.
- **Adjust Cutoff Based on Conditions:** The sales team should adjust the cutoff score based on revenue shortfalls or surpluses and the availability of manpower:
 - **Revenue Shortfall and Manpower Availability:** If there is a shortfall in sales revenue and sufficient manpower, the team can lower the cutoff score to contact more leads, even if this slightly reduces the conversion rate.
 - **On-Track Sales Targets:** When sales targets are on track and manpower is limited, use a higher cutoff score to focus on leads with better conversion potential, improving efficiency and allowing the team to address other tasks.



THANK YOU