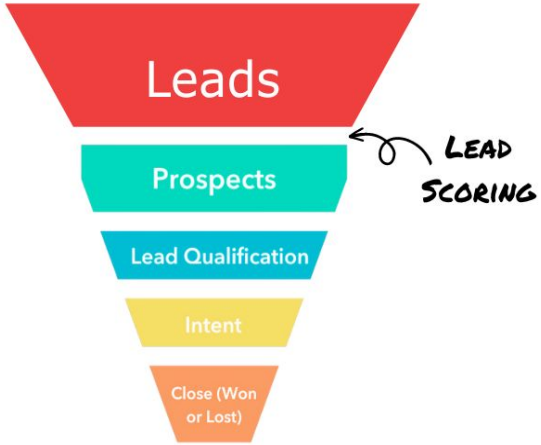
Sales & Mkt Summit: Optimizing Lead Conversion for X Education

Yuleica Ledezma
Data Analyst, Corporate Sales and Marketing

Executive Summary

- **Objective:** Boost lead conversion rates by predicting high-conversion leads using machine learning models.
- **Current Process:** Manual lead scoring based on intuition and limited data, resulting in inefficiency.
- **Goal:** Develop a predictive model that improves the sales **team's focus and efficiency, providing more precise leads,** which will result after implementation *(outside the scope of this project)* to raising the current conversion rate from 38%.



Performance Metrics

•Not only focusing on **Accuracy** (how often the model is correct)

Precision: Fewer false positives, enabling the sales team to focus on higher-quality leads.

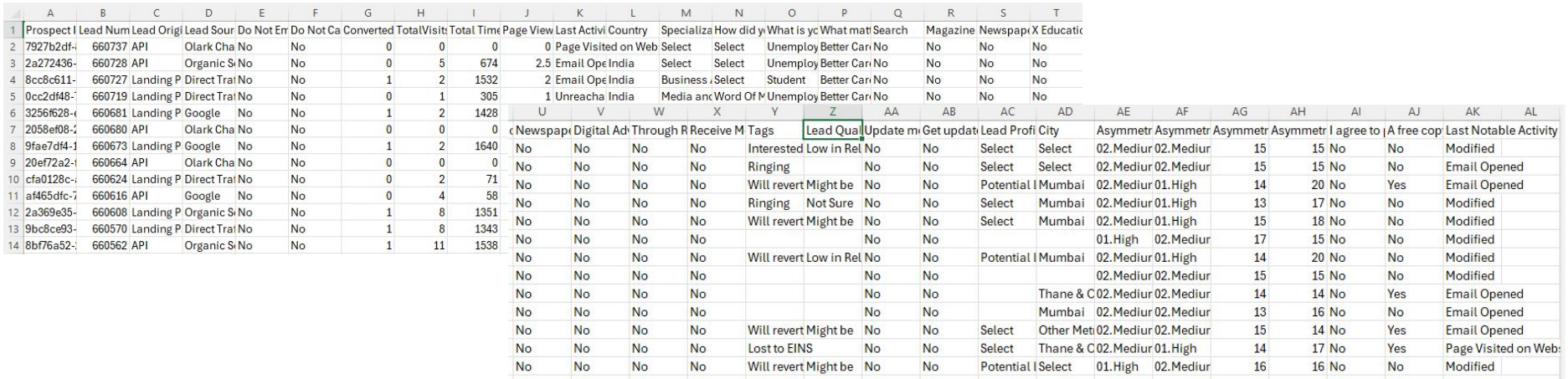
•**Recall:** Captured more potential leads, ensuring no opportunities were missed.

Goal to reach 85% on each

Data Sources

From our partner department [Kaggel Leads Dataset](), they created a dataset that comprises of:

- Website Activity
- Interest Survey/ Program Inquiry response
- Sales interactions
- Manual Scoring based on leads activity (large set of missing values)



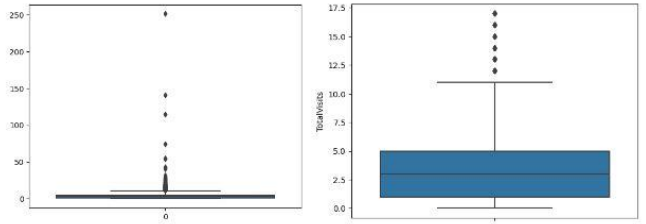
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Prospect I	Lead Num	Lead Origi	Lead Sour	Do Not Em	Do Not Ca	Converted	TotalVisit	Total Time	Page View	Last Activi	Country	Specializa	How did y	What is yc	What matt	Search	Magazine	Newspape	X Educatic
2	7927b2df-	660737	API	Olark Cha	No	No	0	0	0	0	Page Visited on Web		Select	Select	Unemploy	Better Car	No	No	No	No
3	2a272436-	660728	API	Organic S	No	No	0	5	674	2.5	Email Ope	India	Select	Select	Unemploy	Better Car	No	No	No	No
4	8cc8c611-	660727	Landing P	Direct Trat	No	No	1	2	1532	2	Email Ope	India	Business /	Select	Student	Better Car	No	No	No	No
5	0cc2df48-	660719	Landing P	Direct Trat	No	No	0	1	305	1	Unreacha	India	Media and	Word Of M	Unemploy	Better Car	No	No	No	No
6	3256f628-	660681	Landing P	Google	No	No	1	2	1428											
7	2058ef08-	660680	API	Olark Cha	No	No	0	0	0											
8	9fae7df4-1	660673	Landing P	Google	No	No	1	2	1640											
9	20ef72a2-1	660664	API	Olark Cha	No	No	0	0	0											
10	cfa0128c-	660624	Landing P	Direct Trat	No	No	0	2	71											
11	af465dfc-7	660616	API	Google	No	No	0	4	58											
12	2a369e35-	660608	Landing P	Organic S	No	No	1	8	1351											
13	9bc8ce93-	660570	Landing P	Direct Trat	No	No	1	8	1343											
14	8bf76a52-	660562	API	Organic S	No	No	1	11	1538											

	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL
1	Newspape	Digital Ad	Through R	Receive M	Tags	Lead Qual	Update m	Get updat	Lead Profi	City	Asymmetr	Asymmetr	Asymmetr	Asymmetr	I agree to	A free cop	Last Notable Activity	
2	No	No	No	No	Interested	Low in Rel	No	No	Select	Select	02.Mediur	02.Mediur	15	15	No	No	Modified	
3	No	No	No	No	Ringing		No	No	Select	Select	02.Mediur	02.Mediur	15	15	No	No	Email Opened	
4	No	No	No	No	Will revert	Might be	No	No	Potential I	Mumbai	02.Mediur	01.High	14	20	No	Yes	Email Opened	
5	No	No	No	No	Ringing	Not Sure	No	No	Select	Mumbai	02.Mediur	01.High	13	17	No	No	Modified	
6	No	No	No	No	Will revert	Might be	No	No	Select	Mumbai	02.Mediur	01.High	15	18	No	No	Modified	
7	No	No	No	No			No	No			01.High	02.Mediur	17	15	No	No	Modified	
8	No	No	No	No	Will revert	Low in Rel	No	No	Potential I	Mumbai	02.Mediur	01.High	14	20	No	No	Modified	
9	No	No	No	No			No	No			02.Mediur	02.Mediur	15	15	No	No	Modified	
10	No	No	No	No			No	No		Thane & C	02.Mediur	02.Mediur	14	14	No	Yes	Email Opened	
11	No	No	No	No			No	No		Mumbai	02.Mediur	02.Mediur	13	16	No	No	Email Opened	
12	No	No	No	No	Will revert	Might be	No	No	Select	Other Met	02.Mediur	02.Mediur	15	14	No	Yes	Email Opened	
13	No	No	No	No	Lost to EINS		No	No	Select	Thane & C	02.Mediur	01.High	14	17	No	Yes	Page Visited on Web	
14	No	No	No	No	Will revert	Might be	No	No	Potential I	Select	01.High	02.Mediur	16	16	No	No	Modified	

Data Overview and Preprocessing

•Dataset Size: **9,241 rows and 37 columns**, sourced from a Kaggel Leads Dataset.

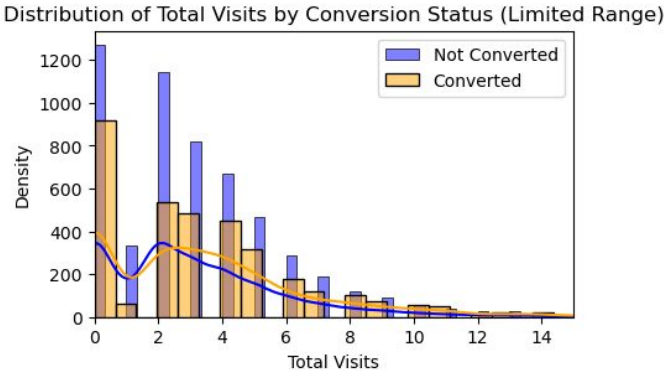
- Numerical (7)
- Categorical (30)



Data Preprocessing: Handle missing values, standardize categorical variables through one-hot encoding, and scale numerical features.

Challenges:

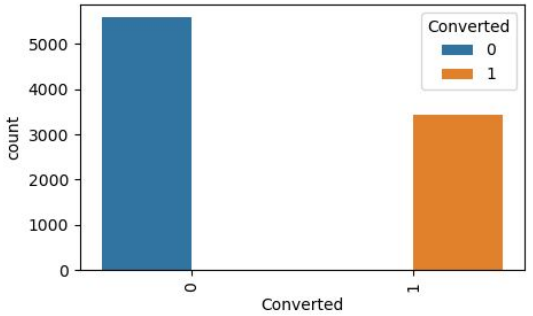
- Large number of null values (~15%, distributed unevenly in different columns)
- Skewness
- Outliers

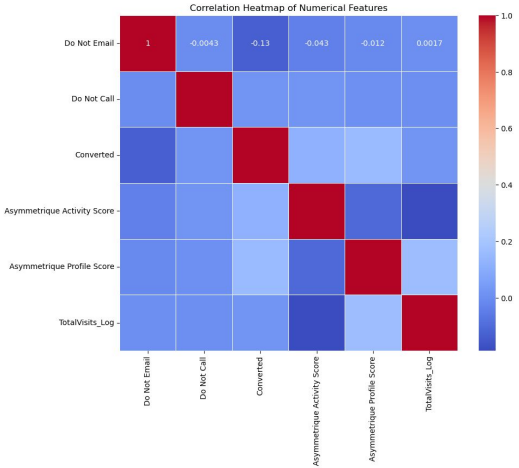


Challenges (cont)

• Significant **class imbalance** (~ 60% not converted vs. ~40% converted), which could affect the models favoring the majority class. For this reason Accuracy is not the only metric, will focus also in Precision and Recall

- Affecting models
- Logistic Regression assumes a balanced class distribution





No strong correlations were found, allowing the model to use a broader set of features

Machine Learning Models Tested

Models Evaluated:

- **Logistic Regression** (Baseline)
  - Logistic Regression with PCA
- **Random Forest**
  - Random Forest with Hyperparameter Tuning (GridSearchCV)
- **Neural Networks**

  Layers: 2 hidden layers

  Neurons:

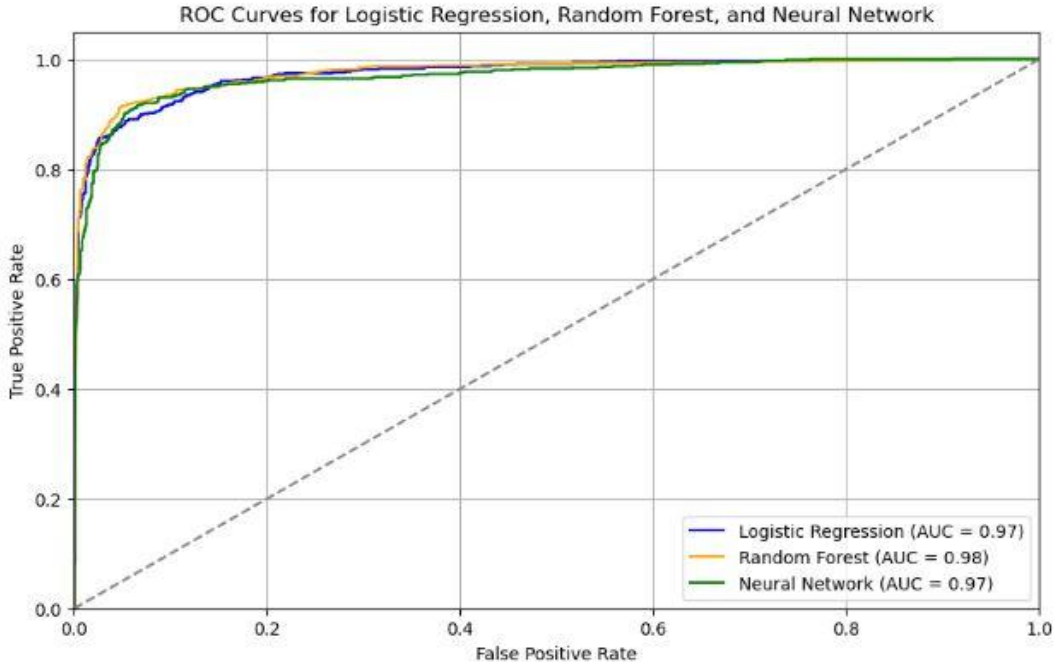
  64 neurons in the first layer

  32 neurons in the second layer

  Activation Functions: ReLU for the hidden layers, Sigmoid for the output layer

  Neural Networks Keras Tuner
- **Performance Metrics**: Accuracy, Precision, Recall, and F1 Score.

Model Evaluation and Results



Model Evaluation and Results

	Log Regression(baseline)	Log Reg PCA	Random Forest	RF Hyperparameter Tuning	Deep Learning	Keras Tuner
Accuracy	.92	.92	.93	.94	.92	.94
Precision	.92	.93	.92	.94	.91	.95
Recall	.87	.87	.90	.90	.90	.89
F1 Score	.89	.90	.91	.92	.91	.91

Recommendation

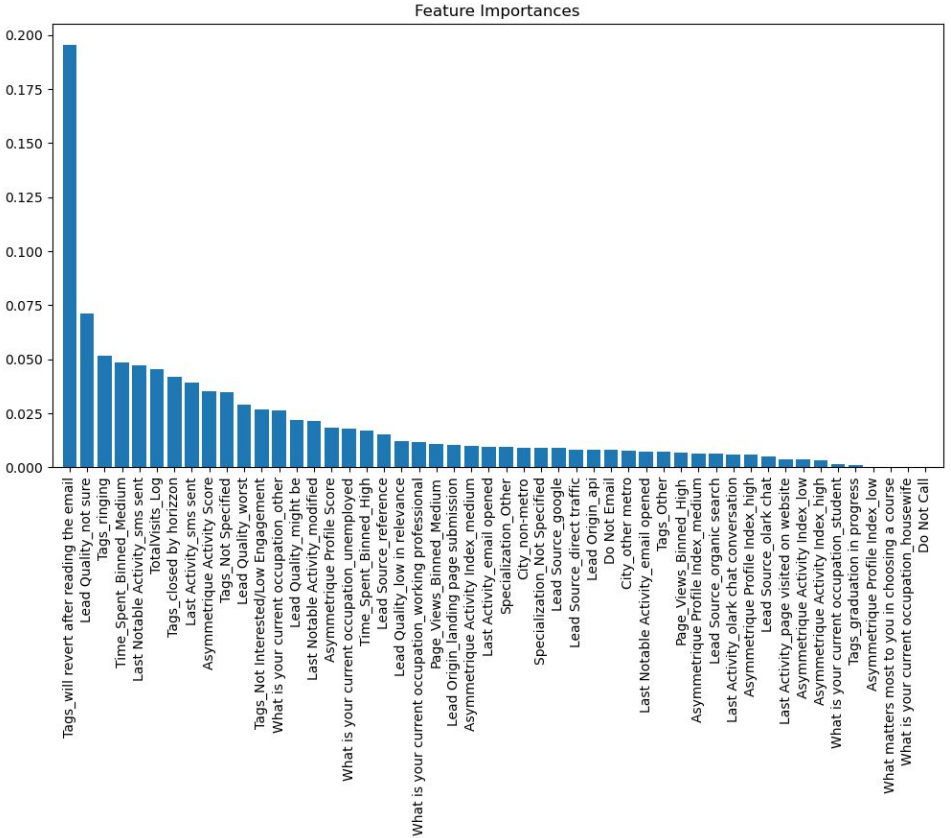
- Recommended Model: **Random Forest with Hyperparameter Tuning.**
  - Achieve a mix between **precision and recall**, ensuring the sales team focuses on the highest-potential leads without missing opportunities.

	Log Regression(baseline)	Log Reg PCA	Random Forest	RF Hyperparameter Tuning	Deep Learning	Keras Tuner
Accuracy	.92	.92	.93	.94	.92	.94
Precision	.92	.93	.92	.94	.91	.95
Recall	.87	.87	.90	.90	.90	.89
F1 Score	.89	.90	.91	.92	.91	.91

Business Impact: Saves time and boosts overall conversions.

Random Forest:

- Works well with **large datasets**.
- Handles **non-linear relationships** effectively.
- **Robust to noise and outliers**.
- Provides **feature importance** insights.
- Handles **class imbalance** and **categorical/numerical data** well.
- **Hyperparameter tuning**, it offers flexibility to achieve high accuracy, precision, and recall.



Key features correlating with conversions: website engagement and last marketing activity.

Conclusion

- Data-driven lead scoring improves efficiency and conversion rates.