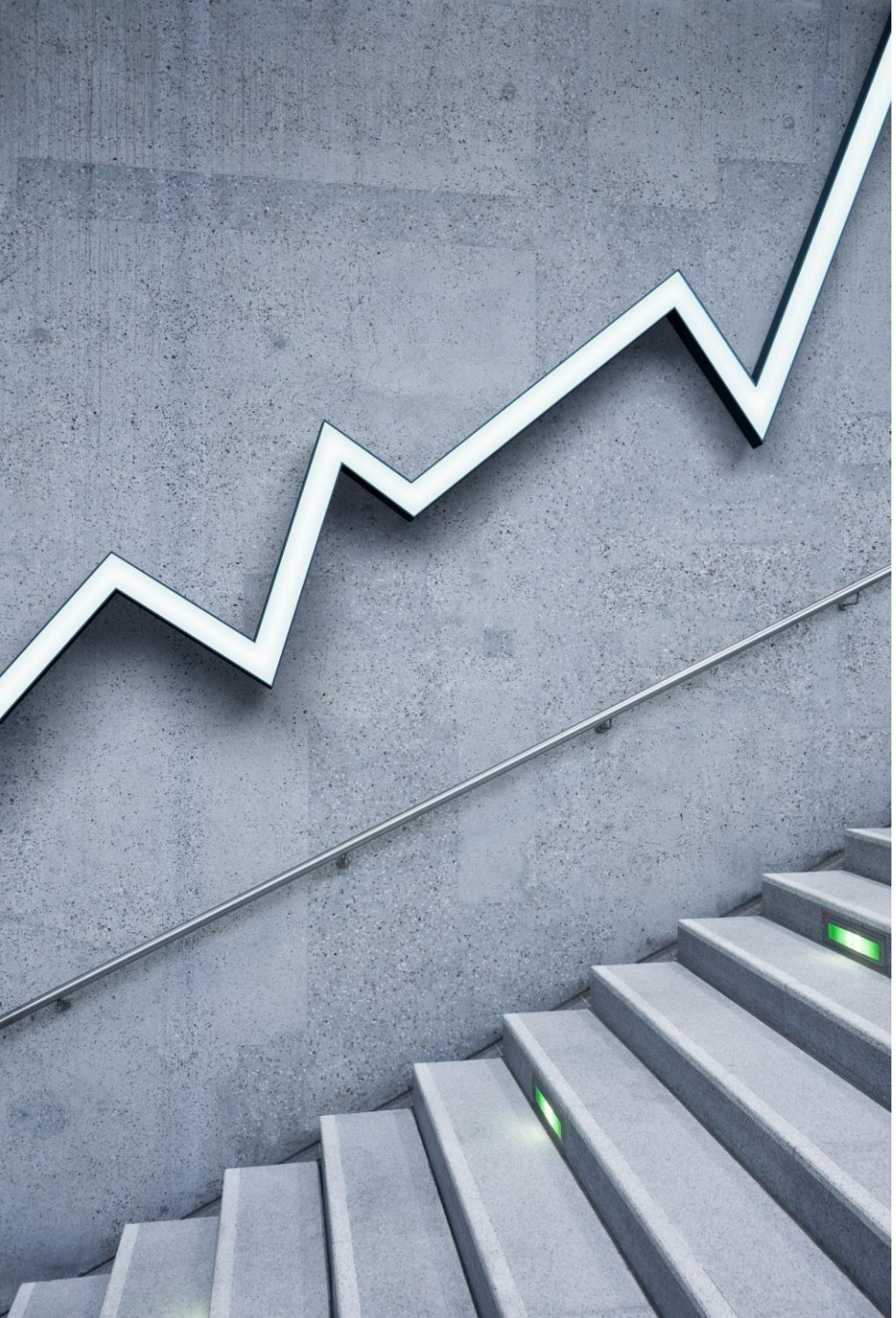


Lead Scoring Case Study

Insha Durwesh

Introduction

Problem Statement:

X Education, an online course provider, faces a low lead conversion rate (~30%) despite acquiring a high volume of leads through various channels like Google and referrals. The company seeks to improve efficiency by identifying high-potential leads, or ‘Hot Leads,’ to focus sales efforts on those most likely to convert. The goal is to increase the conversion rate to around 80% by assigning a lead score based on key attributes such as Lead Source, Total Time Spent on Website, and Last Activity, allowing targeted communication and better resource allocation.

Data Overview:

- 9000+ data points of past leads, with attributes like Lead Source, Website Time, and Last Activity.
- The target variable is 'Converted' (1: Converted, 0: Not Converted).

Objective:

- Develop a logistic regression model that assigns a lead score between 0 and 100 to each lead, indicating the likelihood of conversion. A higher score suggests a 'hot lead' with a strong potential for conversion, while a lower score indicates a 'cold lead' that is less likely to convert.
- Ensure the model is adaptable to future business requirements, allowing X Education to adjust the lead scoring process as needed based on changing sales strategies or market conditions..

Assumptions

- **High Missing Values**: Columns with over 30% missing data were deemed not useful based on the data dictionary and were removed to enhance data quality and model performance.

- **Non-Impactful Geographic Data**: Since X Education operates as an online platform with a global audience, the missing values in the "City" and "Country" columns are not considered impactful, and thus, these columns were dropped.

- **'Select' Values as Missing**: Columns with predominantly 'Select' values, like 'How did you hear about X Education' and 'Lead Profile', were treated as missing data and removed. However, columns like 'Specialization' that have 'Select' but are critical to the analysis were retained.

- **Low Variance Columns**: Variables where a single value overwhelmingly dominates (e.g., 'Do Not Call', 'Search', and others) were excluded since they do not contribute meaningful variation to the model.

- **Imbalanced Responses**: The column 'What matters most to you in choosing a course' was removed due to the significant imbalance in its levels, with one option ('Better Career Prospects') appearing 6528 times and the other options appearing only once or twice.

- **Treatment of 'Select' Values**: 'Select' values were treated as missing, and the count of these occurrences was calculated across relevant columns to decide on their inclusion or exclusion

Data Processing

- **Initial Dataset**:
  9240 rows and 37 columns

- **Duplicates Check**:
  No duplicates were found, ensuring data integrity.

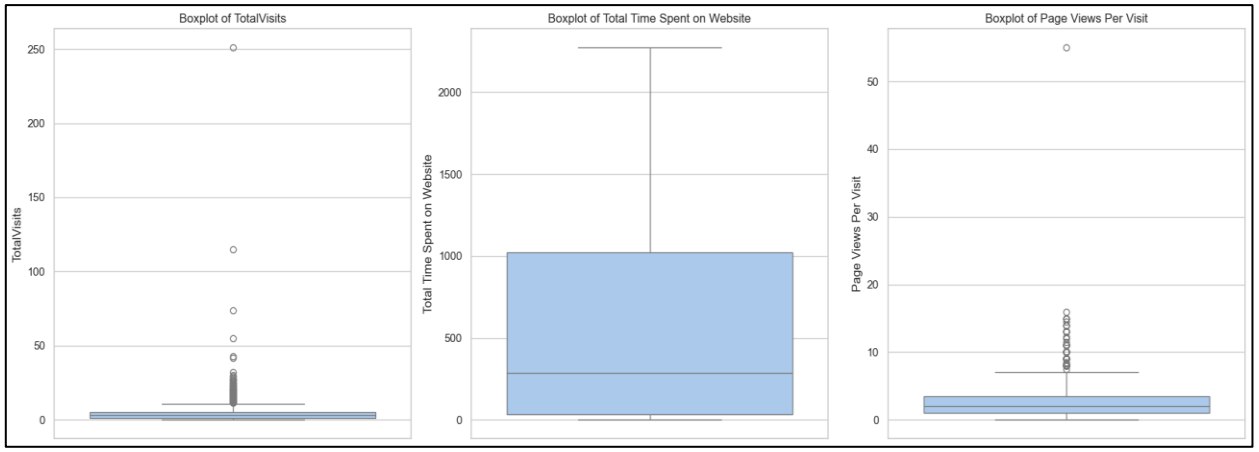
- **Handling Missing Values**:
  - *Dropped Columns:*
    Columns with a high percentage of missing values (>30%) were removed to prevent bias and improve model accuracy.
  - *Dropped Rows:*
    For important features, rows with missing data were removed to maintain the quality of key insights.

- **Low Variance Columns**:
  Columns with little to no variance (e.g., predominantly one value) were excluded because they provide minimal value in predictive modeling and do not help differentiate between leads.

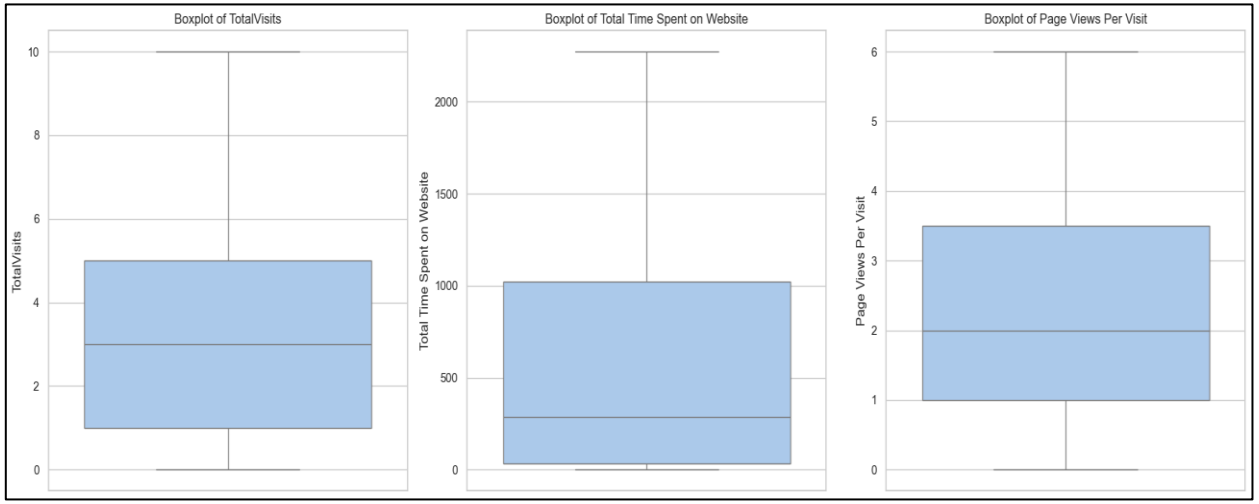
Outlier Treatment

- We used Clipping to remove any outliers from our data frame that may affect our scaling.
- 'TotalVisits' and 'Pages Views Per Visit' had some outliers, which were removed using clipping.

Before Clipping



After Clipping



Preparing Data for Modelling

- To prepare the data for modeling, the next step was to address the categorical variables in the dataset. First, we identified which variables were categorical and then created dummy variables for them.
- This ensured that the model treats these categories as unordered factors without assigning any hierarchical importance.
- We split our dataset into 70% training and 30% testing data.
- We only scaled the numerical variables and not the categorical variables with dummies.
- We used MinMaxScaler, MinMaxScaler transforms features by scaling each feature to a given range, usually between 0 and 1.
- We did not scale y_train or y_test. The target variable 'Converted' is left in its original scale to maintain interpretability and because the scaling of features is intended to ensure that all features are on a comparable range for model training purposes. The target variable itself should reflect the true counts to be predicted by the model.

Model Building Using RFE

- We performed logistic regression on the training data using statsmodels. In statsmodels, we must explicitly fit a constant using sm.add_constant(X) to ensure that the regression line does not pass through the origin by default.
- The dataset contains numerous variables, so we used RFE to select a small set of features from this pool of variables.
- Afterwards, we fitted a logistic regression model on our training data and iterated through 5 models to arrive at the final one. Initially, RFE shortlisted 15 features, but after eliminating the ones with high VIF to address multicollinearity and those with high p-values, we ended up with 11 features.

Insights from Model

- The final model, a Generalized Linear Model (GLM) with a logit link function, has demonstrated strong performance in predicting customer conversions.
- Based on the summary of model coefficients, several key factors significantly impact conversion likelihood, with high statistical significance (p-values < 0.05 for all variables). These include Total Time Spent on Website, Lead Source (Olark Chat, Reference, Welingak Website), and Last Activity (SMS Sent, Unreachable).
- Conversely, factors like Do Not Email, Last Activity (Converted to Lead, Olark Chat Conversation), and current occupation as Student or Unemployed have a negative influence on conversion.

Final Model Summary

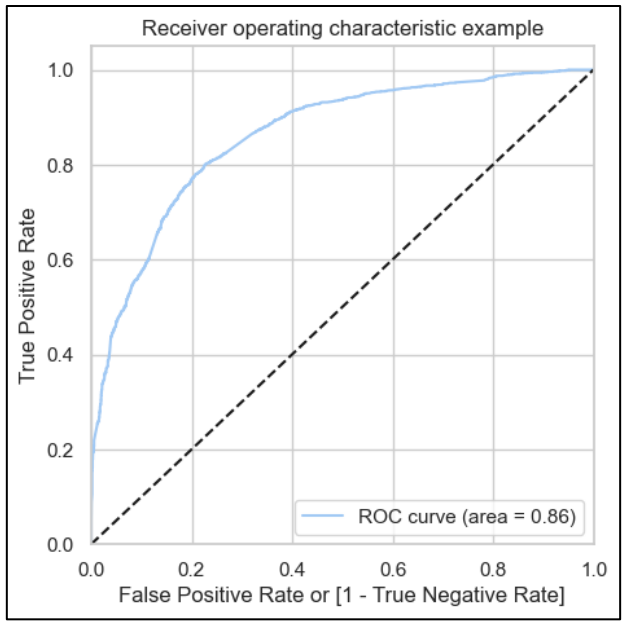
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	4461
Model:	GLM	Df Residuals:	4449
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2069.2
Date:	Tue, 24 Sep 2024	Deviance:	4138.5
Time:	17:51:59	Pearson chi2:	4.48e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3666
Covariance Type:	nonrobust		

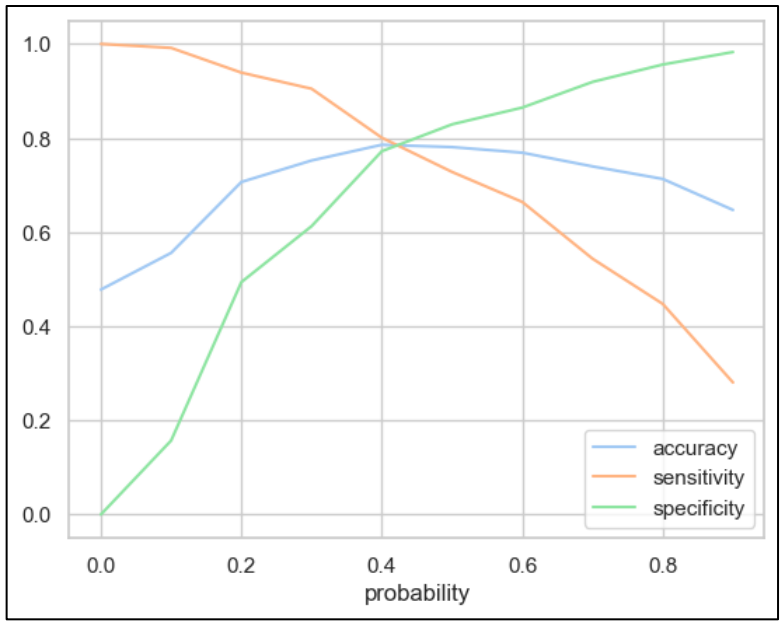
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.7019	0.196	3.590	0.000	0.319	1.085
Total Time Spent on Website	4.3600	0.184	23.708	0.000	4.000	4.720
Lead Source_Olark Chat	1.4232	0.118	12.101	0.000	1.193	1.654
Lead Source_Reference	3.4586	0.229	15.119	0.000	3.010	3.907
Lead Source_Welingak Website	5.4547	0.725	7.526	0.000	4.034	6.875
Do Not Email_Yes	-1.5365	0.192	-8.023	0.000	-1.912	-1.161
Last Activity_Converted to Lead	-1.2336	0.235	-5.255	0.000	-1.694	-0.774
Last Activity_Olark Chat Conversation	-1.3361	0.184	-7.276	0.000	-1.696	-0.976
Last Activity_SMS Sent	1.0519	0.083	12.620	0.000	0.889	1.215
What is your current occupation_Student	-2.5957	0.284	-9.151	0.000	-3.152	-2.040
What is your current occupation_Unemployed	-2.6755	0.191	-14.016	0.000	-3.050	-2.301
Last Notable Activity_Unreachable	2.5532	0.815	3.133	0.002	0.956	4.150

Approach I: Using ROC AUC Curve (Sensitivity and Specificity Tradeoff) to find Optimal Cutoff

The ROC curve has an area of 0.86, indicating a strong model.



The optimal cutoff point by assessing sensitivity and specificity tradeoff is around 0.42.



Evaluation Metrics of Approach I

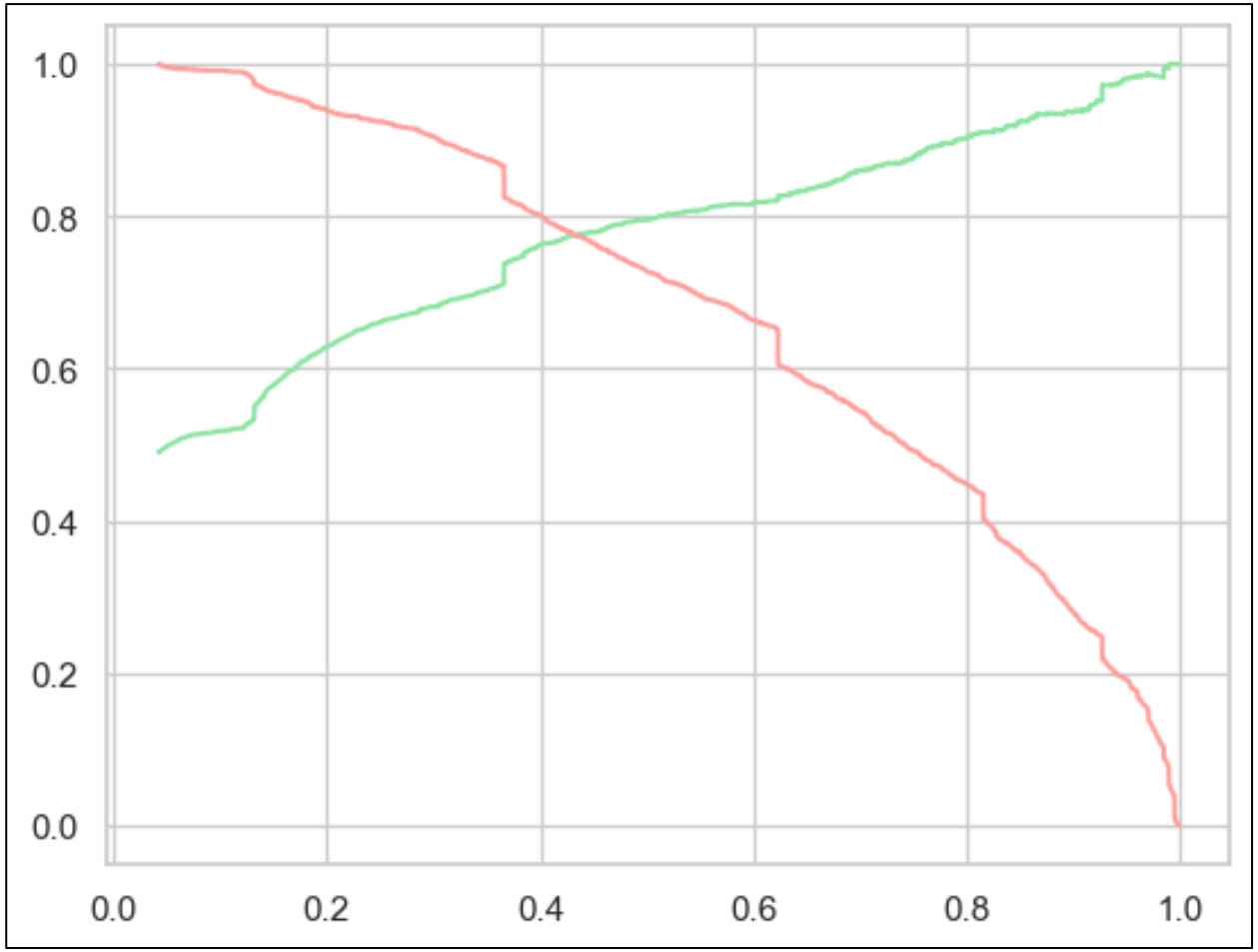
- **Evaluation using ROC AUC Curve:**
  1. **Accuracy**: 79%
  2. **Precision**: 79%
  3. **Recall (Sensitivity)**: 79%
  4. **Specificity**: 80%
  5. **F1-Score**: 79%

- This cutoff provided a balanced trade-off between precision and recall. The model was able to correctly predict **733 true positives (converted customers)** and **784 true negatives (non-converted customers)**. However, it produced **195 false positives** and **200 false negatives**.

Approach II: Using Precision-Recall Curve to find Optimal Cutoff

- Using Precision- Recall Curve, the optimal cutoff we obtained is around 0.44.

Precision-Recall Curve



Evaluation Metrics of Approach II

- **Evaluation using Precision-Recall Curve:**
    1. **Accuracy**: 80%
    2. **Precision**: 81%
    3. **Recall (Sensitivity)**: 77%
    4. **Specificity**: 82%
    5. **F1-Score**: 79%
- Using this cutoff, the model slightly improved in terms of overall accuracy (80%) and precision (81%), but had a marginal decrease in recall (77%). The **trade-off here** is favoring precision over recall, resulting in fewer false positives (175) but a slight increase in false negatives (210).

Conclusion on the Model

- The logistic regression model developed for X Education has proven effective in predicting lead conversion, addressing the company's core challenge of improving its low conversion rate (~30%). By identifying key variables such as Total Time Spent on Website, Lead Source (Olark Chat, Reference, Welingak Website), and Last Activity (SMS Sent, Unreachable), the model provides actionable insights for targeting high-potential leads. Conversely, factors like Do Not Email and certain Last Activities (e.g., Olark Chat Conversation) negatively influence conversion probability, helping the sales team avoid low-priority leads.

- The model offers flexibility depending on the business's priorities. When minimizing false positives is more critical, the Precision-Recall Curve cutoff can be adopted, reducing unnecessary marketing efforts on unlikely conversions. Conversely, the ROC AUC cutoff provides a balanced approach, identifying potential leads while maintaining a solid recall rate.

- In conclusion, the model is well-suited for X Education's objectives, enhancing the lead conversion process by enabling focused sales strategies. With a robust predictive performance, the model can be fine-tuned to align with changing business goals, ultimately increasing the conversion rate and optimizing resource allocation.

Key Takeaways for X Education

- **Focus on Hot Leads**: Prioritize leads with higher scores based on factors like **Total Time Spent on Website** and **Lead Source** (Olark Chat, Reference).

- **Leverage Lead Sources**: Channels like **Welingak Website** and **Olark Chat** show high conversion rates—focus marketing efforts here.

- **Reduce Unnecessary Outreach**: Avoid over-investing in leads marked as **Do Not Email** or those with **low activity engagement** (e.g., Olark Chat Conversation, low website time).

- **Adjust Communication for Occupations**: Tailor strategies for leads in specific occupations (e.g., **Students** and **Unemployed** are less likely to convert).

- **Fine-Tune Sales Strategies**: Use the model's flexibility to adjust lead targeting depending on business goals—maximize conversion during key periods, minimize unnecessary outreach when goals are met.

- **Monitor and Adapt**: Regularly assess lead performance metrics and adjust the lead-scoring model based on changing market conditions and customer behavior trends.

Thank You!

Turning chaos into clarity, one data point at a time.