**Intermediate Progress Report**

Mary Morkos, Alara Kaymak, Laura Li, Camden Bibro

**Summary of EDA Findings, Missing Values, and New Features**

**Missing Value Handling:**

* Dropped Columns: All columns with more than 90% missing values were dropped, including fields like industry metadata, web tracking features, and trial-related dates, as they lacked sufficient coverage for analysis or modeling.
* Dropped Rows: Fully empty rows were removed to clean the dataset at the row level.
* No Imputation on High-Missing Fields: Instead of using forward/backward fill or imputation strategies for columns with minimal utility or inconsistent formatting, we dropped them to ensure only high-integrity features were retained.

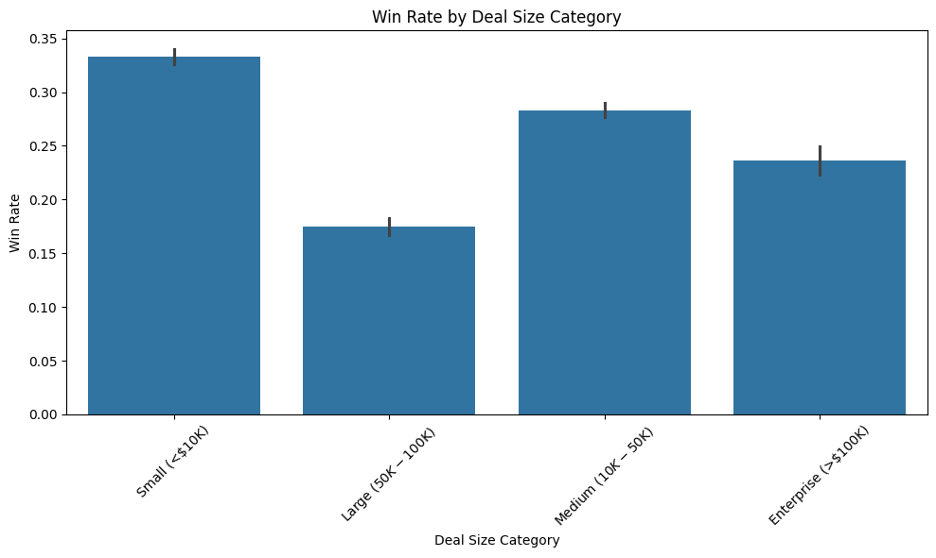
o Initially handled missing values one way during EDA (e.g., dropped high-missing columns, cleaned rows, no imputation)

o Later, different techniques during modeling, like imputation or encoding, depending on what was needed for model performance.

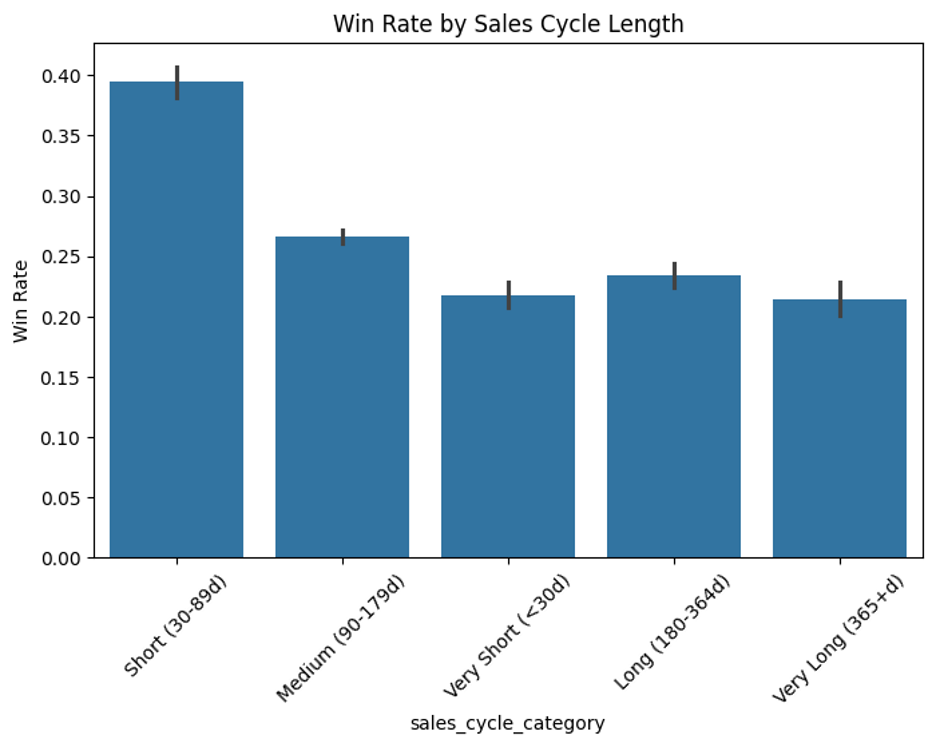
**Feature Engineering:**

* deal\_size\_category: Created from weighted\_amount or fallback amount, categorized into:
  + *Small (< $10K)*
  + *Medium ($10K–$50K)*
  + *Large ($50K–$100K)*
  + *Enterprise (≥ $100K)*This helps us evaluate performance by deal volume vs. value.
* sales\_cycle\_category: Derived from days\_to\_close, binned into:
  + *Very Short (<30 days)*
  + *Short (30–89 days)*
  + *Medium (90–179 days)*
  + *Long (180–364 days)*
  + *Very Long (≥365 days)*This helps analyze how deal duration affects success.
* create\_year\_month: Extracted from create\_date to explore seasonal and trend-based behaviors in deal performance.
* is\_won: A binary outcome flag indicating if deal\_stage == Closed Won. This is the target variable for any downstream classification or win-rate analysis.
* total\_pipeline\_time\_sec: Combined time across multiple deal stages, converted from various hh:mm:ss columns (e.g., BANT, negotiation, contract sent). This was critical for capturing overall time-to-close dynamics.

**Notable EDA Insights:**

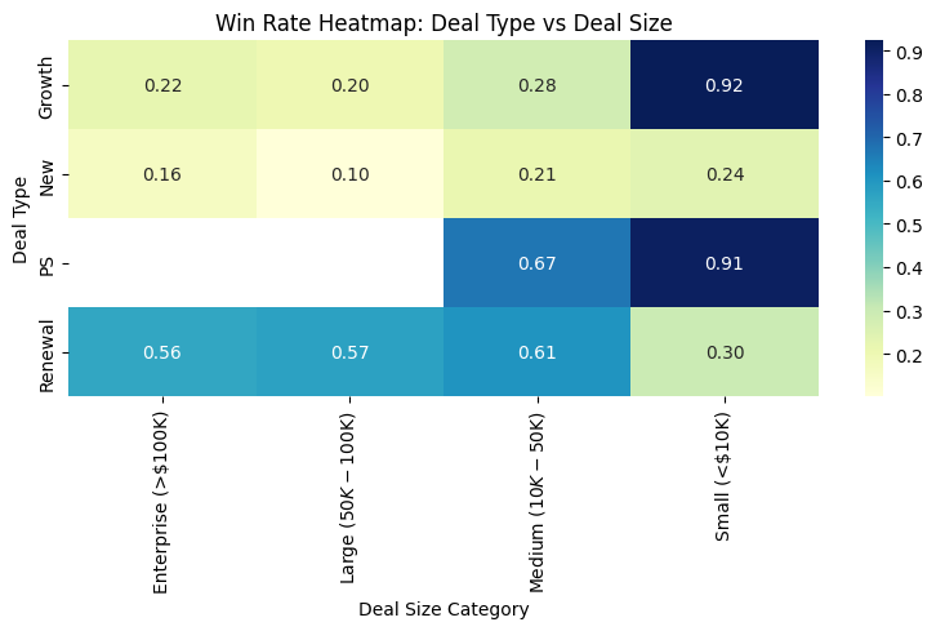


* Smaller Deals Win More: Deals under $10K had noticeably higher win rates. This suggests volume-focused strategies might convert more frequently, though revenue per deal is lower.
* Enterprise Deals Are High Risk, High Reward: Despite lower win rates, they represent significant value and should be managed with strategic precision.



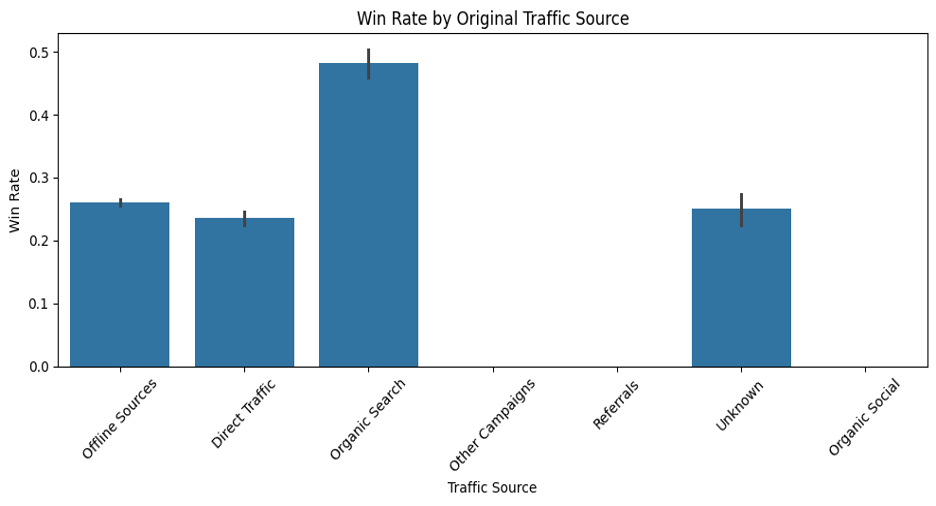
* Shorter Sales Cycles → Higher Success: Most successful deals are closed within 30 to 90 days, indicating that speed and engagement are key success factors.

Smaller deals (<$10K) tend to have the highest win rates, while larger deals have lower win rates. This suggests that smaller deals are generally easier to close, possibly due to lower risk or faster decision-making processes.



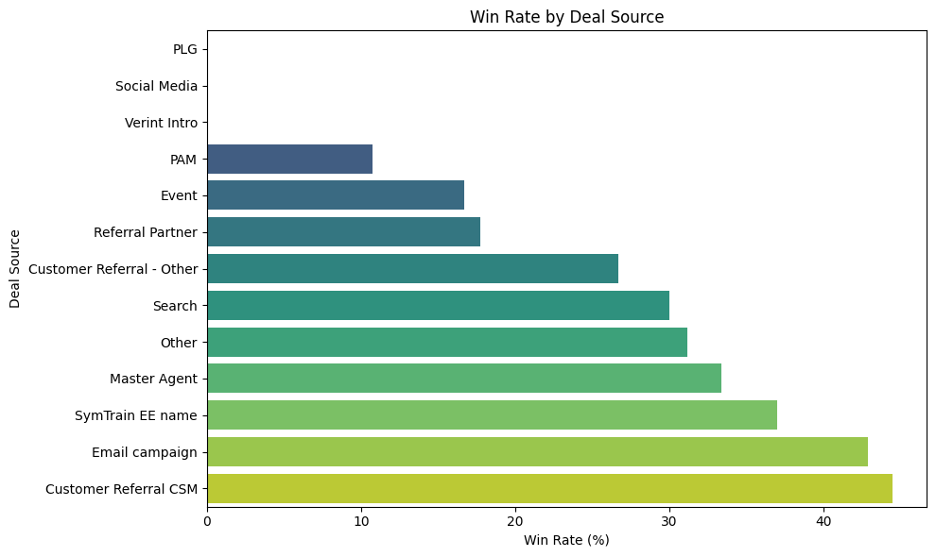
This heatmap breaks the win rate down by both deal type (e.g., New, PS, Renewal, Growth) and deal size. Each cell shows the success rate for that specific combo.

* Professional Services (PS) and Renewals have high win rates, especially for Small and Medium deals.
* New deals have lower success across all sizes.
* Growth deals do well at the smallest deal sizes, but drop off elsewhere.



* Professional Services (PS) deals have the highest win rate (~86%), suggesting that PS offerings are highly successful, likely due to existing client trust or tailored needs.
* Renewals also perform well (~50%), indicating strong retention and customer satisfaction.
* Growth deals sit in the middle (~40%).
* New deals have the lowest win rate (~18%), highlighting the challenge of acquiring brand-new customers compared to expanding or retaining current ones.

This insight supports prioritizing upsell and service-based strategies over cold acquisition when optimizing win rates.



**Customer Referrals & Email Campaigns Lead in Conversion:** Deals sourced through Customer Referral – CSM and Email Campaigns showed the highest win rates (~45% and ~43%), followed by SymTrain EE Name and Master Agent channels.

This suggests that relationship-driven and targeted outreach channels perform better in driving successful deal closures, while sources like PAM and Social Media had notably lower conversion rates.

**Deal Stage ↔ Deal Probability:**

The deal probability field aligns closely with deal stage, scoring 1.0 for "Closed Won", 0.0 for "Closed Lost", and intermediate values for open deals. While not central to this EDA, it could be useful as a confidence signal for downstream modeling or dashboards.

**Tickets Data EDA, Missing Value Handling, and Feature Engineering:**

**Missing Value Handling:**

This dataset consists of 79 tickets from a call service company, each representing a service or implementation process tracked through multiple project stages. Upon inspection, the dataset contains 46 columns, many of which are timestamps for different project milestones, as well as text-based notes, identifiers, training indicators, and metadata.

Dropped Columns: Columns with more than 90% missing values were removed from the analysis. These included "Ticket Tags", "Category", "Last CES survey rating", "Planning Phase", and "Was the sym QAed?". These features had minimal usable data and offered little analytical value for either exploratory data analysis (EDA) or predictive modeling.

Datetime Standardization: All date-related fields were converted into standard datetime format to allow for time-difference calculations and the extraction of time-based features such as weekday, hour, and month. No imputation was performed on datetime fields at this stage; missing dates were retained to study stage dropout and ticket progression trends.

Stage Progression: The dataset contains multiple columns starting with "Stage Date -", each corresponding to a phase in the ticket lifecycle:

* "Stage Date - Project Initiation"
* "Stage Date - Planning Phase"
* "Stage Date - Execution"
* "Stage Date - Monitoring and Control Phase"
* "Stage Date - Closure Phase"
* "Stage Date - Project Launch"
* "Stage Date - Converted Won"

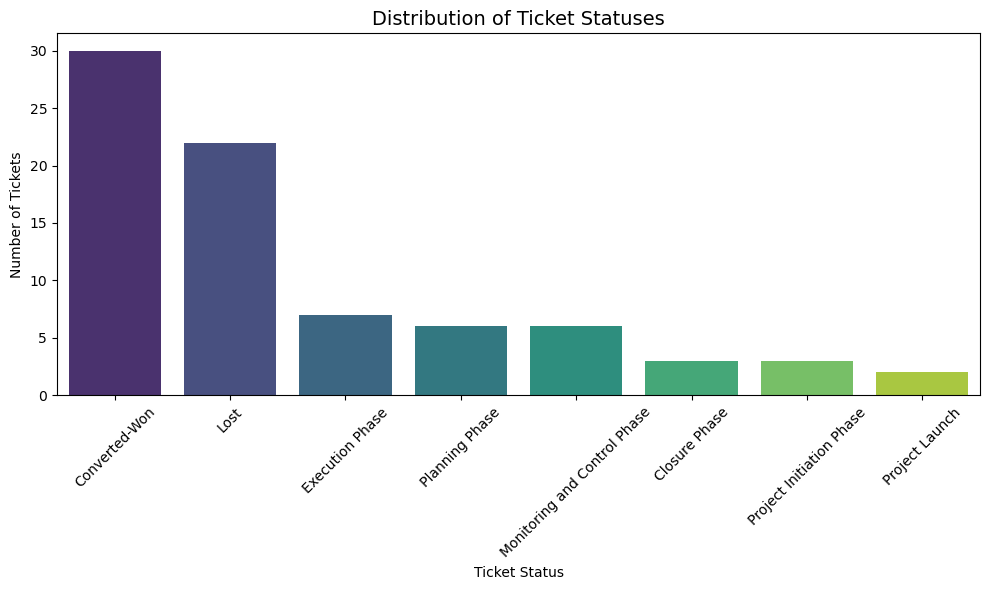
Each of these represents a milestone that a ticket may reach during its lifecycle. Tickets with non-null values for a stage are considered to have progressed to that phase. Using this logic, we created a feature called "num\_stages\_completed" to quantify each ticket's overall progression.

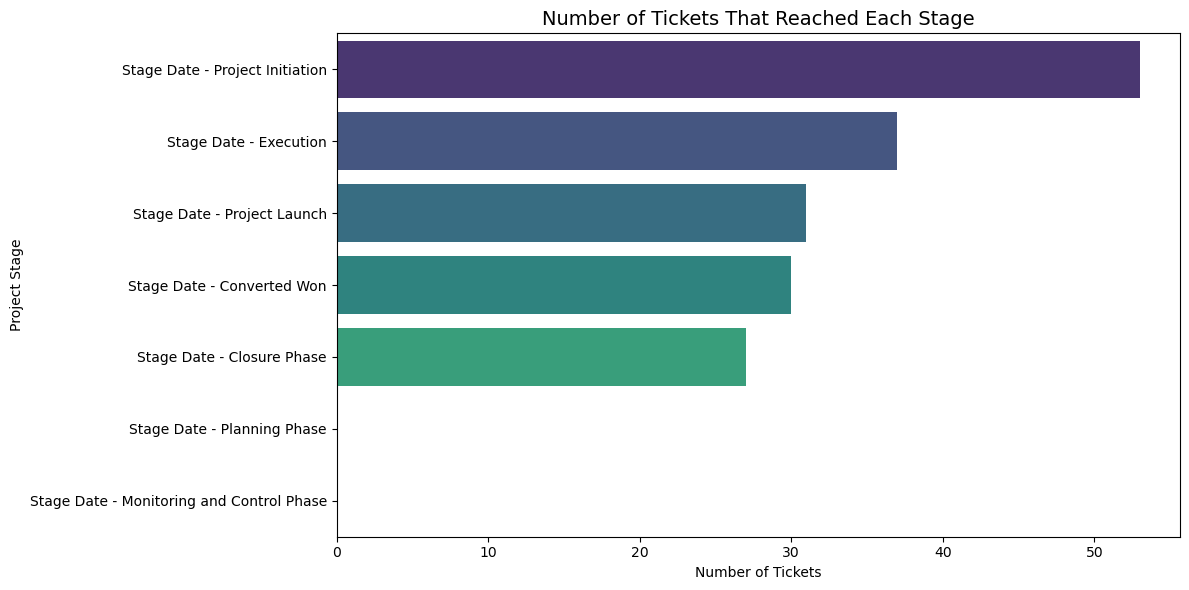
No imputation was performed on stage data. Instead, we computed which transitions were present for each ticket and used this to identify stage bottlenecks and average durations.

Feature Engineering:

The following features were engineered to support modeling and enhance EDA:

* "num\_stages\_completed": Count of non-null project stage milestones reached by the ticket.
* "converted\_won": A binary outcome flag indicating if the ticket reached the "Converted Won" stage. This serves as our target variable for downstream classification.
* "total\_duration\_days": Total time in days from ticket creation ("Create date") to closure ("Close date").
* "num\_transitions\_with\_duration": The number of measurable stage-to-stage time transitions within a ticket.
* Time Features:  
  "create\_month": Month ticket was created (1 to 12)  
  + "create\_dayofweek": Day of the week ticket was created (0 = Monday)
  + "create\_hour": Hour of the day ticket was created
  + "is\_weekend": Binary flag for tickets opened on a Saturday or Sunday
* Stage Transition Durations: We calculated the time in days between consecutive project stages where possible. These durations provide insight into the velocity of tickets through the pipeline. Examples:
  + "Stage Date - Project Initiation" to "Stage Date - Execution"
  + "Stage Date - Closure Phase" to "Stage Date - Project Launch"

Notable EDA Insights:

Stage Participation:

* Out of 79 tickets, only a small subset progressed through all six defined project stages.
* The most common stages reached were "Project Initiation" and "Execution". Very few tickets reached "Converted Won" or "Monitoring and Control".
* This suggests either a high dropout rate or a streamlined pipeline for most tickets.

Ticket Status Distribution:

* The most frequently occurring ticket status was "In Progress", followed by "Closed" and "On Hold".
* This indicates that many tickets remain active or are paused rather than concluded.

Stage Timing:

* Average transition durations were longest between the Execution and Closure phases, suggesting that the final phase of implementation or validation may be the most complex or delayed.
* The fastest transitions occurred early in the lifecycle, particularly between Project Initiation and Execution.

Modeling Considerations:

The engineered dataset is suitable for classification modeling aimed at predicting whether a ticket will reach the "Converted Won" stage. Our immediate next steps include:

* Encoding categorical variables (e.g., "Ticket status", "Pipeline")
* Handling remaining missing values through imputation (median for duration, mode for categorical)
* Evaluating baseline classification models such as Logistic Regression and Random Forest

Special focus will be placed on evaluating how time-based features, stage duration, and training participation affect the likelihood of successful ticket conversion. A future enhancement may include incorporating textual ticket notes using NLP for sentiment or theme extraction.

This EDA phase has laid the groundwork for targeted modeling and insights into operational bottlenecks, project performance, and time-to-success for call service tickets.

**Missing Value Handling for Models**

1.For the data used to train model for deal outcome prediction (deals+company, tickets were dropped due to >70% missing):

1. Removed columns with over 70% missing
2. Date columns (a combination of median, forward & backward fill, taking most frequent date, depending on if missing <5%,5%-30%, or over 30%)
3. Time columns: median imputation
4. Categorical: Mode imputation, for industry columns used cross-filling, for columns with >10 unique categories filled NA with unknown
5. Numeric: <30% missing used median imputation, columns like (pageview, sessions) defaulted to 0, for 30%-70% missing used MICE (random forest regressor, max iter 5, fallback to median)

2. For the data used to train clustering algorithm:

1. Removed columns with over 40% missing
2. Removed rows with remaining missing values

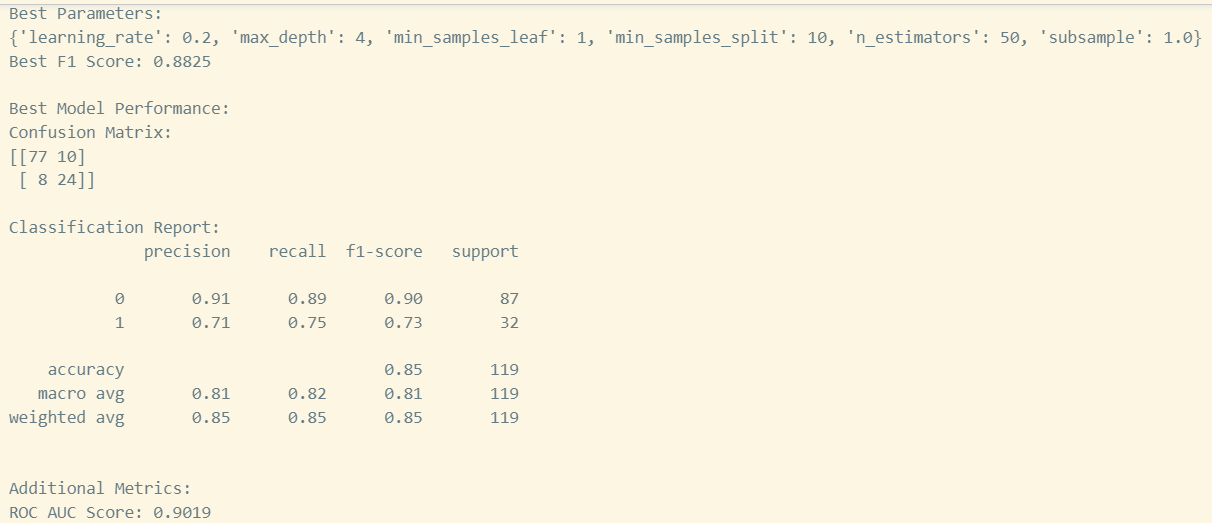
**Status of data cleaning, integration, and preprocessing**

* Two merged versions were created, one is between deals+company+tickets (left join on deals), the other is a left join on based on ticket
* Each dataset is cleaned/missing data handled separately for EDA (left join on deals, left join on tickets) and modeling (left join on deals, company dataset on its own)

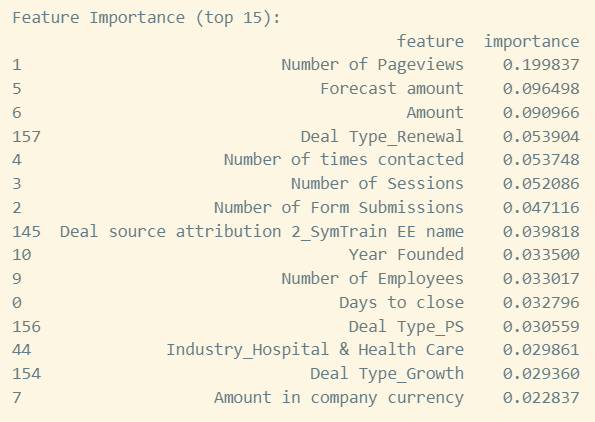
**Baseline models tested, initial results, and planned improvements**

Deal Outcome Prediction Model - Gradient Boosting Classifier:

1. Target Variable (Is Closed Won) Distribution:
   * Not Closed Won: 73.36% (lost), Closed Won: 26.64%
   * Use SMOTE resampling, oversampled minority
2. Feature Selection:
   * Business Metrics**:** Days to close,Deal source attribution, Original Traffic Source, Pipeline, Deal Type
   * Financial Indicators: Forecast amount, Amount, Amount in company currency, Annual Revenue
   * Company Characteristics: ICP Fit Level, Industry, Primary Industry, Number of Employees, Year Founded, Segmentation
   * Engagement: Number of Pageviews**,** Number of Form Submissions**,** Number of Sessions**,** Number of times contacted
   * One hot encoding for categorical variables, exclusion of leaky features
3. Hyperparameters and Performance



1. Top Features:



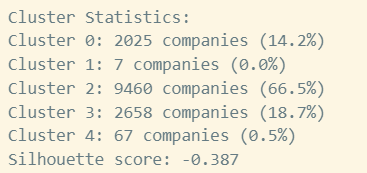
Companies data based clustering algorithm (PCA+K-means clustering):

Categorical Feature Processing:

* Transformed Industry data (171 unique values) into 10 business sectors such as Technology, Finance, Healthcare
* Categorized Web Technologies (13,446 unique values) into 8 technology groups including Analytics, Cloud Services, Marketing
* Grouped geographical data into 4 major regions: North America, Europe, Asia Pacific, and Latin America
* Selected and scaled numeric variables using StandardScaler
* Created dummy variables for the processed categorical features
* Filtered out rare categories (< 1% representation) to avoid sparse data issues
* Combined both feature types into a unified feature matrix with 28 dimensions
* Applied PCA to reduce the 28-dimensional feature space
* Retained sufficient components to explain 92% of the variance in the data

Clustering Results:

* Applied K-means clustering with k=5 clusters
* Evaluated cluster quality using silhouette score and distribution analysis
* Results: PCA explained variance: 0.92



**Key focus areas before the final submission**

1. Feature engineering to create interaction terms
2. Explore alternative oversampling techniques
3. Consider other models
4. Investigate why pageviews are such a strong predictor
5. Implement regularization techniques
6. Experiment with different feature selection
7. Adjust algorithm, hyperparameter and also data pre-handling steps for clustering algorithm, currently using PCA and K-means and clusters are quite imbalanced