**Intermediate Progress Report:**

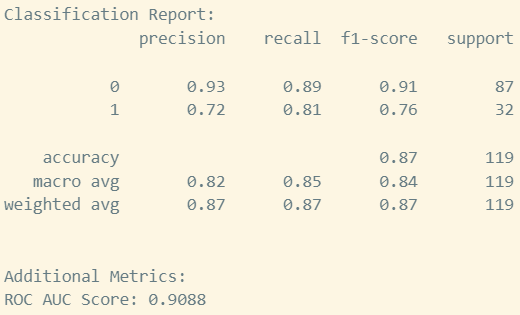
•**Summary of EDA findings, missing values, and any new features created**

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 <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: median imputation, columns like (pageview, sessions) defaulted to 0, for 30%-70% missing used MICE (random forest regressor, max iter 5, fallback to median)

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

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

Deal Outcome Prediction Mode- Gradient Boosting Classifier:

1. Target Variable (Is Closed Won) Distribution:
   * Not Closed Won: 73.36%, 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:
   * n\_estimators: 100
   * learning\_rate: 0.1
   * max\_depth: 3
   * Random state: 42
4. Performance:
   * Cross-validation results: mean f1 score 0.8721 (±0.0519)
   * 
5. Top Features: Number of Pageviews (22.64%), Amount (7.28%), Forecast amount (6.98%), Amount in company currency (5.62%), Deal Type\_Renewal (5.41%), Deal source attribution 2\_SymTrain EE name (5.29%), Deal Type\_PS (4.72%), Number of times contacted (4.53%), Number of Sessions (4.00%)

•**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