Lead Conversion Prediction – Project Summary

By Umar Farooq Bhat

Objective

The objective of this project was to predict which leads are most likely to convert for a higher education company. This would help the marketing and sales teams focus their efforts on high-potential leads, improving overall efficiency and conversion rates.

Data Understanding

The dataset consisted of 9240 records and 37 columns, containing a mix of numerical, categorical, and text-based features. The target variable was Converted – indicating whether a lead successfully enrolled (1) or not (0). Initial exploration revealed missing values, duplicate entries, and several irrelevant or low-variance columns which were addressed during preprocessing.

Data Cleaning and EDA

Missing values were handled either by imputation or dropping where necessary. Duplicate rows were removed, and columns with excessive nulls were eliminated.
Key insights included:
• Leads from Olark Chat, Google, and Direct Traffic had the highest conversion rates.
• Tags like “Will revert after reading the email” and “Interested in Next Batch” indicated high intent.
• Specializations such as HR and Marketing correlated with higher conversions.
• Total Time Spent on Website and Page Views were strong indicators of conversion likelihood.

Feature Engineering

Categorical variables were encoded using dummy variables. Numerical features were scaled for logistic regression. Multicollinearity was addressed using Variance Inflation Factor (VIF), and features were selected based on statistical significance (p-values).

Model Building

Two models were trained:
• Logistic Regression (with VIF and p-value-based feature selection)
• Random Forest Classifier (for comparison and feature importance)

Train/test split: 80/20 (stratified).
Threshold tuning and cross-validation were used for robustness.

Model Evaluation

• Accuracy: 96%
• Precision: 96.3%
• Recall: 94.1%
• F1-score: 95.2%
• ROC AUC: 0.986

Random Forest also performed comparably, validating results.

Business Recommendations

1. Prioritize leads with “Interested” tags or engagement history.
2. Focus budget on high-performing sources like Google and Olark Chat.
3. Improve underperforming lead sources (e.g., Facebook).
4. Encourage longer time spent on site and better landing page design.
5. Deploy model for automated lead scoring.

Conclusion

The final model is highly accurate, interpretable, and ready for deployment. With an AUC of 0.986, it supports strategic decision-making in both marketing and sales, ensuring higher conversion efficiency and better customer targeting