Click Through Rate Predictions

Agenda

1. Objective
2. Background
3. Reasons for Selecting the 3 Models (Classification Method)
4. Evaluation Metrics Chosen
5. Key Finding (for all 3 Models: for both the Balanced & Imbalanced Dataset)
6. Recommendations for the Best Model
7. Risks & Implications of the Errors Associated with Recommending of The Best Model
8. Conclusion
9. Appendix:

    a. Data Sources

    b. Data Preparation

    c. Data Methodology

Objective

To prepare the Machine Learning Model (ML) to predict whether the audience will click on an ad or not.

For that:

a) Identify the Classification Methods (3 - Chosen for this Project's ML Model Building)

b) Justification for Choosing the 3 Classification Methods to meet the objective

c) Recommend the One Model (among the 3) for final deploying to the Business Team

d) Explain the Risks Associated with with that Final Recommended Model for Deploying

Background

1. I, as Data Scientist (Hypothetical), working in a Marketing Company have been tasked to prepare ML model for predicting whether the audience will click an ad or not
2. For that Click Through Rate Prediction (Important Metric used for evaluating Ad- Performance) is used
3. CTR not just helps in prediction of Final Click but often helps in answering market related questions, for example: Which ad to use and whom to target it for , Where to launch the ad (which platform - mobile/desktop/app/web page etc)

The Three Different Models

The 3 classification methods used for building the ML Models are as follows :

Logistic Regression

Decision Trees

Random Forests

Furthermore these models will be trained and tested again after careful applications of model simplification techniques like feature creation and selection, to improve model predictability.

Reasons For Choosing Logistic Regression (LR)

1) Simplicity and Interpretability: LR provides a straightforward interpretation of models coefficient, making it suitable when interpretability is crucial
2) Linear Relationship: LR assumes a linear relationship between the input features and the log - odds of the target class
3) Binary Classification: LR is specifically designed for binary classification problems and works well when the goal is to predict one of the two classes

Reasons For Choosing Decision Tree (DT)

1. Non Linear Relationship: DT can capture complex non linear relationships b/w features & the target variable. They are effective in scenarios where the underlying relationship isn't linear
2. Feature- Interactions: DT naturally handle feature interactions, where the combination of multiple features has a different effect than the individual features themself
3. Easy Interpretation: DT can be easily visualised & interpreted, providing insights into how the model makes decisions

Reasons For Choosing Random Forest (RF)

1. Robustness and Generalisation: RF are an assembly of DT, which helps improve the model's generalisation and robustness by reducing overfitting
2. Handling High Dimensional Data: RF can handle datasets with large no.'s of features without overfitting. They are effective in high dimensional scenarios
3. Feature Importance: RF can rank the importance of features based on their contribution to the model performance. This information can help to identify the most relevant features

Evaluation Metrics

The evaluation metrics for click-through rate (CTR) prediction provide valuable insights.

Accuracy measures the overall correctness of the model's predictions.

Precision focuses on the proportion of correctly predicted positive instances (CTR) among all predicted positives.

Recall evaluates the proportion of correctly predicted positive instances among all actual positives.

The area under the Receiver Operating Characteristic (ROC) curve assesses the model's ability to distinguish between positive and negative instances. Higher values indicate better performance in all these metrics.

Key Findings – (Dummy Classifier)

Model	Accuracy_train	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean	CrossVal1	CrossVal2	CrossVal3	CrossVal4	CrossVal5
Dummy classifier	0.830340433	0	0	0.829866667	0	0	0.830340434	0.830285714	0.830357143	0.830357143	0.830357143	0.830345025

1. The "Dummy classifier" model, as shown in the table, demonstrates high accuracy on both the training (approx. 0.83) and testing (approx. 0.82) datasets but fails to correctly identify any positive instances, resulting in a recall and precision of zero

2. The cross-validation results indicate consistent performance with a mean value (approx. 0.83) similar to the individual folds.

Key Findings – LR with Imbalanced Data Before RFE

Model	Accuracy_train	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean	CrossVal1	CrossVal2	CrossVal3	CrossVal4	CrossVal5
logistic regres - imbal	0.831040443	0.573134328	0.01616706	0.830766667	0.604651163	0.015282132	0.83052615	0.830142857	0.83	0.831142857	0.830857143	0.830487892

1) The "Logistic regression - imbalance predict" model, as depicted in the table, shows relatively high accuracy on both the training and testing datasets
2) However, it exhibits imbalanced prediction performance, with a higher recall on the training set (57.31%) compared to the precision (1.62%)
3) The cross-validation results indicate consistent performance with a mean value (approx. 0.83) similar to the individual folds.

Key Findings – LR with Balanced Data Before RFE

Model	Accuracy_train	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean	CrossVal1	CrossVal2	CrossVal3	CrossVal4	CrossVal5
logistic regres - balance	0.681887013	0.68604098	0.670298836	0.67905324	0.685308018	0.663176475	0.67639775	0.671513379	0.672875887	0.678812648	0.679716068	0.679070768

1) The "Logistic regression - balance predict" model, as presented in the table, demonstrates lower accuracy (approx 0.67-0.68) compared to the previous model (approx. 0.82-0.83) on both the training and testing datasets
2) It shows improved balance in predicting positive instances, with higher recall (approx. 0.68) and precision values (0.67 - 0.66) for both the training and testing sets
3) The cross-validation results indicate consistent performance across the folds, with a mean value (approx. 0.67) close to the individual fold values

Key Findings – LR with RFE

Model	Accuracy_train	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean	CrossVal1	CrossVal2	CrossVal3	CrossVal4	CrossVal5
logistic reg with feature	0.694655241	0.681250301	0.731219445	0.694250381	0.682693395	0.72684238	0.694672424	0.697496343	0.69580555	0.691546569	0.692493009	0.69602065

1) The "Logistic regression with feature engineering" model, as depicted in the table, demonstrates improved accuracy (approx. 0.69) compared to the previous models on both the training and testing datasets
2) It shows a balanced performance in terms of recall (approx 0.68) and precision (approx 0.72-0.73) for both the training and testing sets
3) The cross-validation results indicate consistent performance across the folds, with a mean value (approx 0.69) close to the individual fold values
4) The inclusion of feature engineering appears to have positively influenced the model's performance

Key Findings – Basic DT - Imbalanced Data

Model	Accuracy_train	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean	CrossVal1	CrossVal2	CrossVal3	CrossVal4	CrossVal5
decision tree basic - imbalan	0.976442521	0.989376974	0.870495116	0.758533333	0.290933959	0.291731975	0.757682238	0.759857143	0.758142857	0.754	0.759785714	0.756625473

1) The "Decision tree basic - imbalance predict" model, as shown in the table, achieves very high accuracy on the training dataset
2) However, it exhibits imbalanced prediction performance, with a high recall on the training set (98.94%) but a low recall on the testing set (29.09%)
3) The precision values are also relatively low
4) The cross-validation results indicate consistent performance across the folds, but the model's performance on the testing data suggests that it may not generalize well to unseen instances.

Key Findings – Basic DT - Balanced Data

Model	Accuracy_train	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean	CrossVal1	CrossVal2	CrossVal3	CrossVal4	CrossVal5
decision tree basic - balance	0.982542632	0.990550024	0.974368243	0.792800932	0.796580572	0.786897741	0.784024205	0.786715994	0.781200258	0.780339858	0.785631319	0.786233599

1) The "Decision tree basic - balance predict" model, as presented in the table, achieves high accuracy on both the training and testing datasets
2) It demonstrates a balanced performance in terms of recall and precision for both the training and testing sets
3) The cross-validation results indicate consistent performance across the folds, with a mean value close to the individual fold values
4) This model, with balanced prediction, appears to generalize better to unseen instances compared to the previous decision tree model with imbalanced prediction

Key Findings – DT - Imbalanced Data- With RFE

Model	Accuracy_train	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean	CrossVal1	CrossVal2	CrossVal3	CrossVal4	CrossVal5
Decision tree with feature er	0.835440506	0.598023064	0.091697541	0.834966667	0.6	0.089929467	0.835326197	0.836642857	0.836928571	0.834	0.835285714	0.833773841

1) The "Decision tree with feature engineering - Imbalanced Dataset" model, as shown in the table, achieves relatively high accuracy on both the training and testing datasets
2) However, it demonstrates imbalanced prediction performance with higher recall values compared to precision values for both the training and testing sets
3) The cross-validation results indicate consistent performance across the folds, with a mean value close to the individual fold values
4) The inclusion of feature engineering appears to have improved the model's performance compared to the basic decision tree model, but imbalanced predictions persist

Key Findings – DT - Balanced Data- With RFE

Model	Accuracy_train	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean	CrossVal1	CrossVal2	CrossVal3	CrossVal4	CrossVal5
Decision tree with feature en	0.782828283	0.774155859	0.798423191	0.760077893	0.755453185	0.769727605	0.757842474	0.758453067	0.757496236	0.758313616	0.760808776	0.754140675

1) The "Decision tree with feature engineering - Balanced Data" model, as depicted in the table, achieves relatively high accuracy on both the training and testing datasets
2) It demonstrates a balanced performance in terms of recall and precision for both the training and testing sets
3) The cross-validation results indicate consistent performance across the folds, with a mean value close to the individual fold values
4) This model, trained on balanced data, shows improved generalization and balanced prediction performance compared to the previous decision tree models.

Key Findings – Basic RF - Imbalanced Data

Model	Accuracy_train	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean	CrossVal1	CrossVal2	CrossVal3	CrossVal4	CrossVal5
Random forest basic - imbala	0.976413949	0.968821641	0.889609296	0.8081	0.378126166	0.198471787	0.808525819	0.809642857	0.809357143	0.804285714	0.812	0.807343382

1) The "Random forest basic - imbalance predict" model, as shown in the table, achieves high accuracy on the training dataset
2) However, it exhibits imbalanced prediction performance, with a higher recall on the training set (96.88%) compared to the precision (88.96%)
3) The recall on the testing set is relatively low (37.81%), indicating difficulty in correctly identifying positive instances The precision on the testing set is also low (19.85%)
4) The cross-validation results indicate consistent performance across the folds, but the imbalanced prediction performance suggests potential challenges in generalizing to unseen instances and accurately identifying positive instances

Key Findings – Basic RF - Balanced Data

Model	Accuracy_train	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean	CrossVal1	CrossVal2	CrossVal3	CrossVal4	CrossVal5
Random forest basic - balanc	0.982499613	0.982975497	0.981994078	0.830322011	0.836375516	0.821679304	0.825391897	0.826765895	0.825080663	0.823488922	0.824822542	0.826801463

1) The "Random forest basic - balance predict" model, as presented in the table, achieves high accuracy on both the training and testing datasets
2) It demonstrates a balanced performance in terms of recall and precision for both the training and testing sets, indicating a good ability to identify positive instances accurately
3) The cross-validation results indicate consistent performance across the folds, with a mean value close to the individual fold values
4) This model, trained on balanced data, shows improved generalization and balanced prediction performance compared to the random forest model with imbalanced prediction

Key Findings – RF - Imbalanced Data - RFE

Model	Accuracy_train	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean	CrossVal1	CrossVal2	CrossVal3	CrossVal4	CrossVal5
Random forest with RFE - imb	0.976142516	0.966965593	0.889777703	0.810366667	0.390735898	0.204937304	0.810711569	0.81	0.810785714	0.808142857	0.814785714	0.80984356

1) The "Random forest with RFE - imbalance predict" model, as depicted in the table, achieves high accuracy on the training dataset
2) However, it exhibits imbalanced prediction performance, with a higher recall on the training set (96.70%) compared to the precision (88.98%)
3) The recall on the testing set is relatively low (39.07%), indicating challenges in correctly identifying positive instances
4) The precision on the testing set is also low (20.49%)
5) The cross-validation results indicate consistent performance across the folds, but the imbalanced prediction performance suggests potential difficulties in generalizing to unseen instances and accurately identifying positive instances
6) The inclusion of Recursive Feature Elimination (RFE) does not seem to significantly improve the model's imbalanced prediction performance

Key Findings – RF - Balanced Data - RFE

Model	Accuracy_train	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean	CrossVal1	CrossVal2	CrossVal3	CrossVal4	CrossVal5
Random forest with RFE - bal	0.981957565	0.982956996	0.980909592	0.827912953	0.833380896	0.820074618	0.822320299	0.823281425	0.824134222	0.818971822	0.822757582	0.822456442

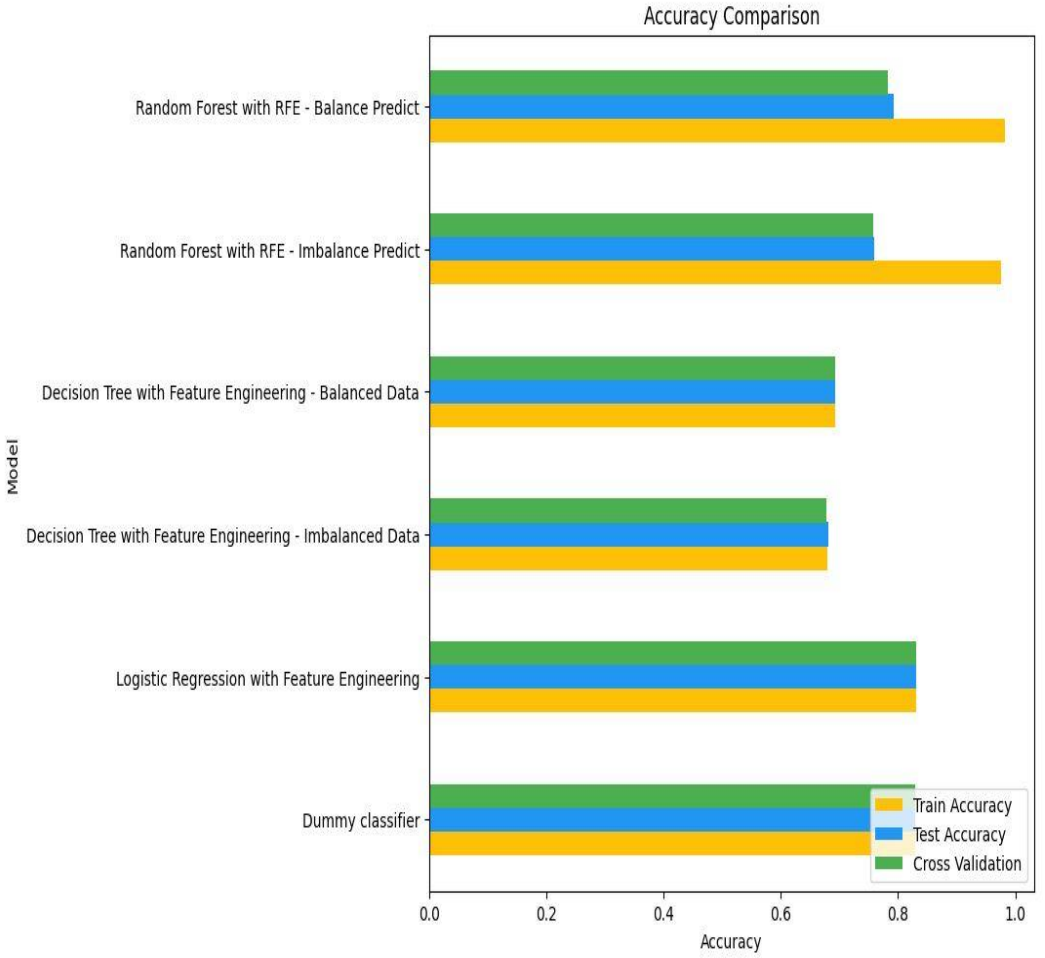
1) The "Random forest with RFE - balance predict" model, as shown in the table, achieves high accuracy on both the training and testing datasets
2) It demonstrates a balanced performance in terms of recall and precision for both the training and testing sets, indicating a good ability to identify positive instances accurately
3) The cross-validation results indicate consistent performance across the folds, with a mean value close to the individual fold values
4) This model, trained on balanced data and with the inclusion of Recursive Feature Elimination (RFE), shows improved generalization and balanced prediction performance compared to the random forest model with imbalanced prediction

Summarising The Key Models

Model	Accuracy_trai	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean
Dummy classifier	0.830340433	0	0	0.829866667	0	0	0.830340434
logistic reg with feature engine- Balanced	0.694655241	0.681250301	0.731219445	0.694250381	0.682693395	0.72684238	0.694672424
Decision tree with feature engine - Imbalanced Dataset	0.835440506	0.598023064	0.091697541	0.834966667	0.6	0.089929467	0.835326197
Decision tree with feature engine - Balanced Data	0.782828283	0.774155859	0.798423191	0.760077893	0.755453185	0.769727605	0.757842474
Random forest with RFE - imbalance predict	0.976142516	0.966965593	0.889777703	0.810366667	0.390735898	0.204937304	0.810711569
Random forest with RFE - balance predict	0.981957565	0.982956996	0.980909592	0.827912953	0.833380896	0.820074618	0.822320299

The Best Model

After applying appropriate evaluation metrics to assess model performance thoroughly, the Random Forest Model after feature engineering (Balanced) seems to perform the best across all metrics



Reason for Recommending the Random Forest as the Best Model

The "Random forest with RFE - balance predict" model appears to be the best choice.

Here's the justification:

1) High Accuracy: The "Random forest with RFE - balance predict" model achieves a high accuracy of 0.9819 on the training dataset, indicating a good overall fit to the data
2) Balanced Performance: The model demonstrates high recall (0.9829) and precision (0.9809) values on the training dataset, indicating a balanced ability to identify positive instances accurately while minimizing false positives
3) Good Performance on Testing Dataset: The model maintains a high recall (0.8334) and precision (0.8201) on the testing dataset, suggesting its ability to generalize well and perform consistently on unseen instances
4) Cross-Validation Mean: The cross-validation mean of 0.8223 indicates consistent performance across multiple folds, reinforcing the model's stability and generalization ability
5) Overall Consistency: The model exhibits consistent performance across different evaluation metrics, as indicated by the values in the table. Consistency in performance further strengthens the model's reliability

Risks Associated

While random forests (after RFE - Balanced Data) offer high accuracy for click-through rate (CTR) prediction, they can be computationally expensive and require substantial memory.

Additionally, the model's complexity makes it challenging to interpret and explain the decision-making process.

Please consider the trade-offs between performance and interpretability before adopting this, as random forests is not a highly interpretable model in terms of understanding the decision making but it is also the best performing model on this dataset.

Implications of Errors

1) The "Random forest with RFE - balance predict" model achieves high accuracy on the training set (98.20%), but performance on unseen data is unknown
2) It shows balanced recall (83.34%) and precision (82.01%) on the testing set, but there is a slight difference between them
3) The cross-validation mean (82.33%) indicates consistent performance across folds, but generalization to unseen data should be assessed
4) Misclassifications could have implications, with false negatives leading to missed opportunities and false positives causing unnecessary actions or resource allocation
5) Random forest models' complexity makes interpretation difficult and identifying biases or errors challenging.

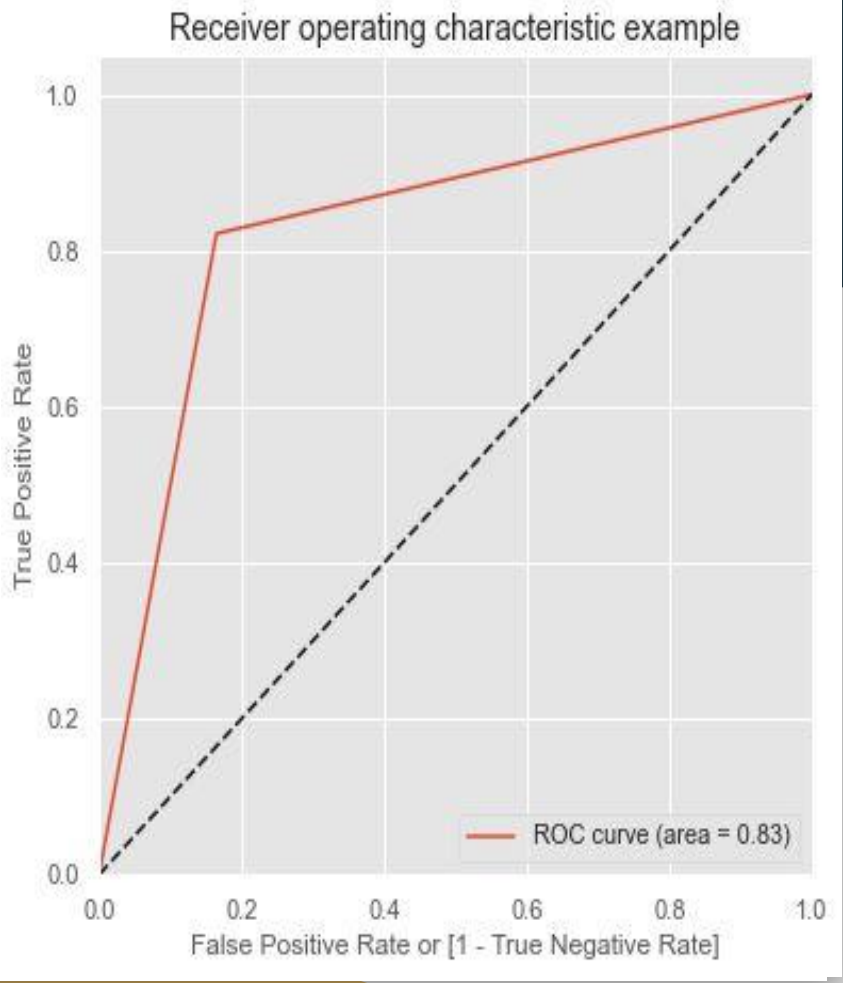
Implications of Errors

5) False positives (Type I errors) occur when the model incorrectly predicts a click when it should not have, potentially leading to wasted resources on irrelevant ads.

6) False negatives (Type II errors) happen when the model fails to predict a click that would have occurred, resulting in missed opportunities for conversions or revenue. Balancing these errors is crucial to optimize the performance of the CTR prediction model.

Conclusion

1) "Random forest with RFE - balance predict" model comes out to be the best model in comparison to all the others in terms of the overall evaluation metrics i.e. Accuracy, Recall and Precision
2) Here. for this model, AUC-ROC is 0.83 (near to 1): implying its ability to strongly distinguish and rank the TPR higher than the negative i.e. FPR



Appendix: Data Sources

Main Data File contains all the information related to the ad, which platform it was shown, whether it was clicked or not

Data Dictionary contains all information for all the features present in the Main Data File

Appendix: Data Preparation

1) Visualised the given dataset through their suitable graphs to gain a better understanding of it
2) Checked for any NA values and any Outliers values (Accordingly dealt with it)
3) Used both Hot Encoding and Label Encoding for the columns where it was better suitable for
4) Checked whether dataset is imbalanced (Accordingly dealt with it)
5) Then ML Models were made for both imbalanced and balanced dataset

Note: Preferred Hot Encoding: where categorical values in a categorical column were not many in terms of their uniqueness. Although Hot Encoding would have been used for categorical columns where unique values were large (keeping in view no relationship between each value in the a particular column. However Label Encoding (LE) was done on them to save our System from getting crashed while running. To use LE over them, proper ordinal relationship was considered in terms of their frequency (value_counts) of a value in that particular categorical column

Appendix: Data Methodology

Different stages/phases that has been followed for this project is mentioned below:

1) Data Collection through Data Sources
2) Data Preparation: Data Analysis through suitable charts/graphs, Treating NA values/ Outliers, Doing Hot and Label Encoding wherever required, and Treating the imbalanced data set
3) Model Building: Choosing the 3 Suitable Model for our Project (predict whether audience will click an ad or not – Classification problem), Building & then Comparing the prediction of all the 3 models (for both imbalanced & balanced data) through Evaluation Metrics (Accuracy, Recall, and Precision) and, AUC ROC Curve
4) Finally Recommending the Best Model for Deploying to the Business Team

Thank You