Shopper Behavior Analysis

Andres Machado, Jackson Small, Sarah Taha, and Thomas Tibbetts

Introduction

Data Profile

- Online Shoppers Purchasing Intention from UCI Machine Learning Repository (*Sakar & Kastro, 2018*).
  - **Disclaimer:** Original retailer anonymized and collection details were undisclosed for privacy reasons; results represent patterns within this dataset only

- Each row describes a visit by a user to an online retailer's webpage.
  - Records Google Analytics information such as duration spent on different pages, bounce rate of entry page, operating system and region.
  - Data was taken over a one-year period (exact dates unknown) of shopping.

- Dataset Dimensions: 12,330 rows and 18 columns.

Feature Overview:

- **Session Structure** 🗂️: Tracking user navigation on different areas of site and their duration
  - Features: *Administrative, Administrative Duration, Informational, Informational Duration, Product Related, and Product Related Information*.
- **Engagement Signals** 📈: Capturing user interaction through measures of different rates
  - Features: *Bounce Rate, Exit Rate, and Page Value.*
- **User Context** 💻: Describes who the user is and how they arrived
  - Features: *Operating System, Browser, Region, Traffic Type, and Visitor Type*
- **Temporal** 🗓️: Representing when the session occurred
  - Features: *Weekend, Month, Special Day*

Objective

Response Column: Revenue (True or False):

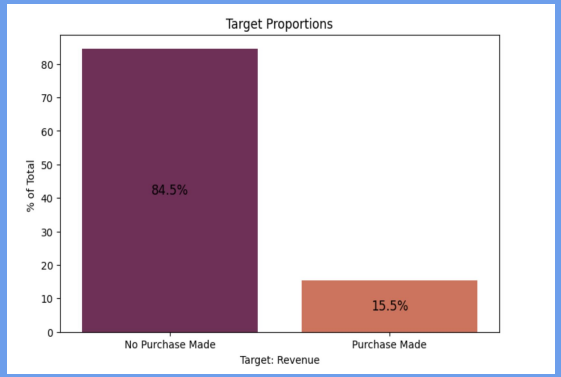
- Use Google Analytics data to predict if a visit will result in a purchase (earning revenue) or not.

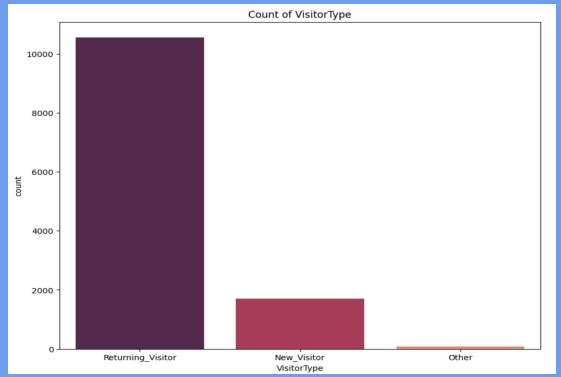
- Compare various classification algorithms with hypertuned parameters:
  - K Nearest Neighbors (KNN)
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Trees and Random Forests

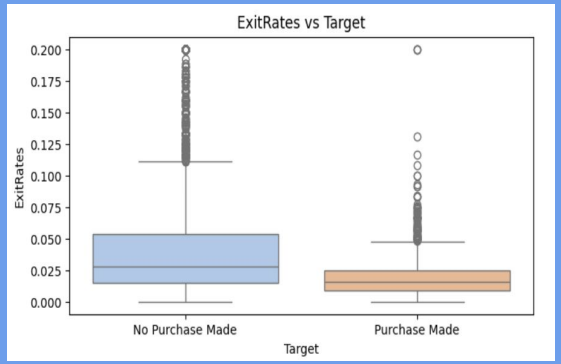
Exploratory Data Analysis

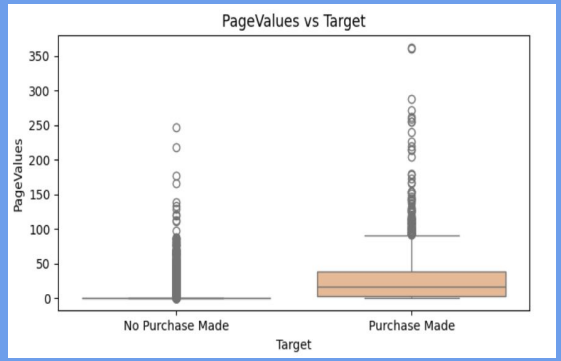
Initial Insights

- Dataset clearly imbalanced shown by the Target (Revenue) proportions.
- More returning users than new users.
- Users who made purchase were more likely to go into pages that lead to a purchase than non-purchasing users.
- Non-purchasing users sporadically exit site pages throughout their sessions more than purchasing users.

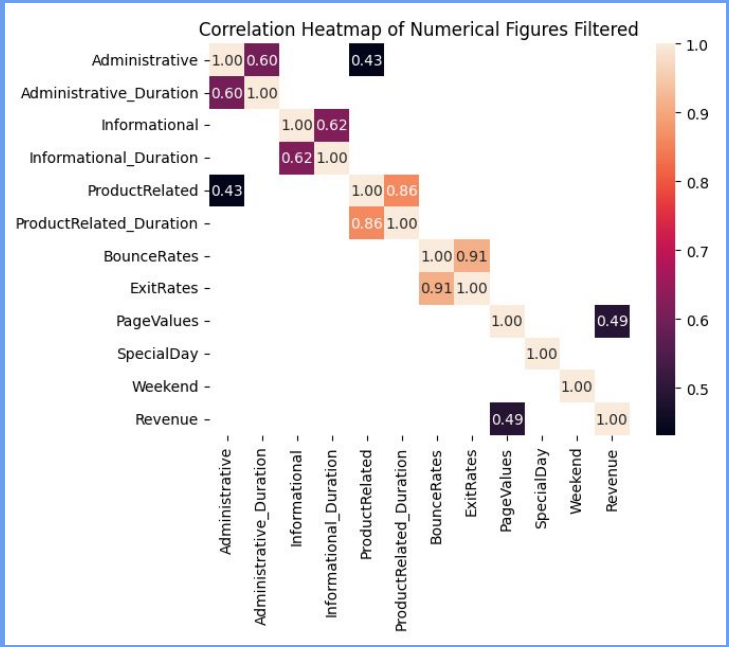




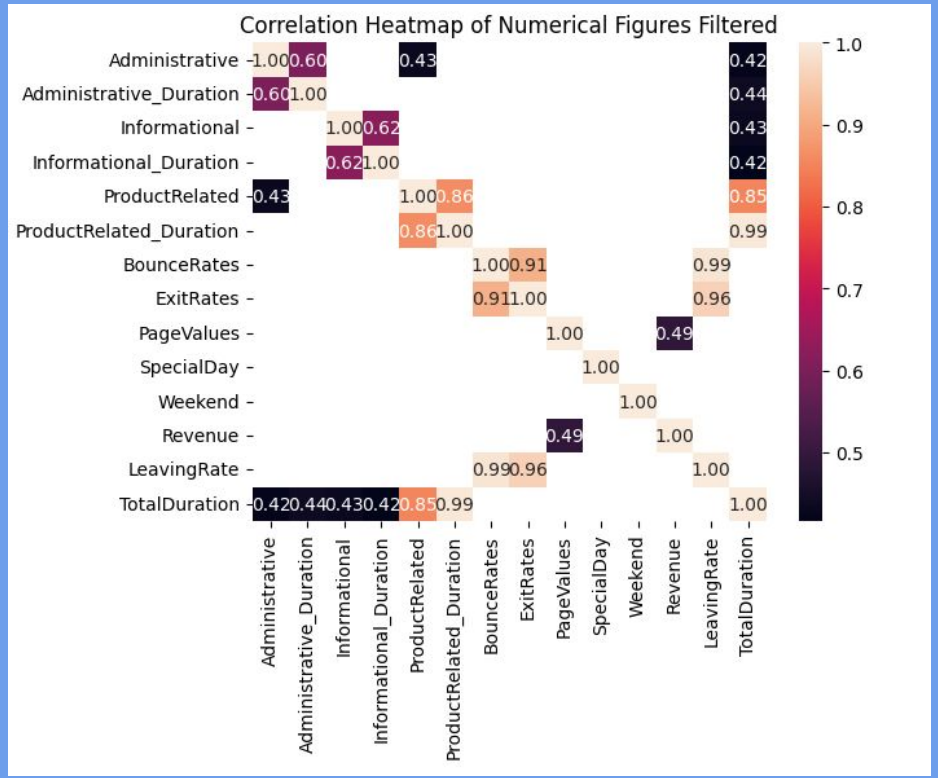




Correlation Matrix

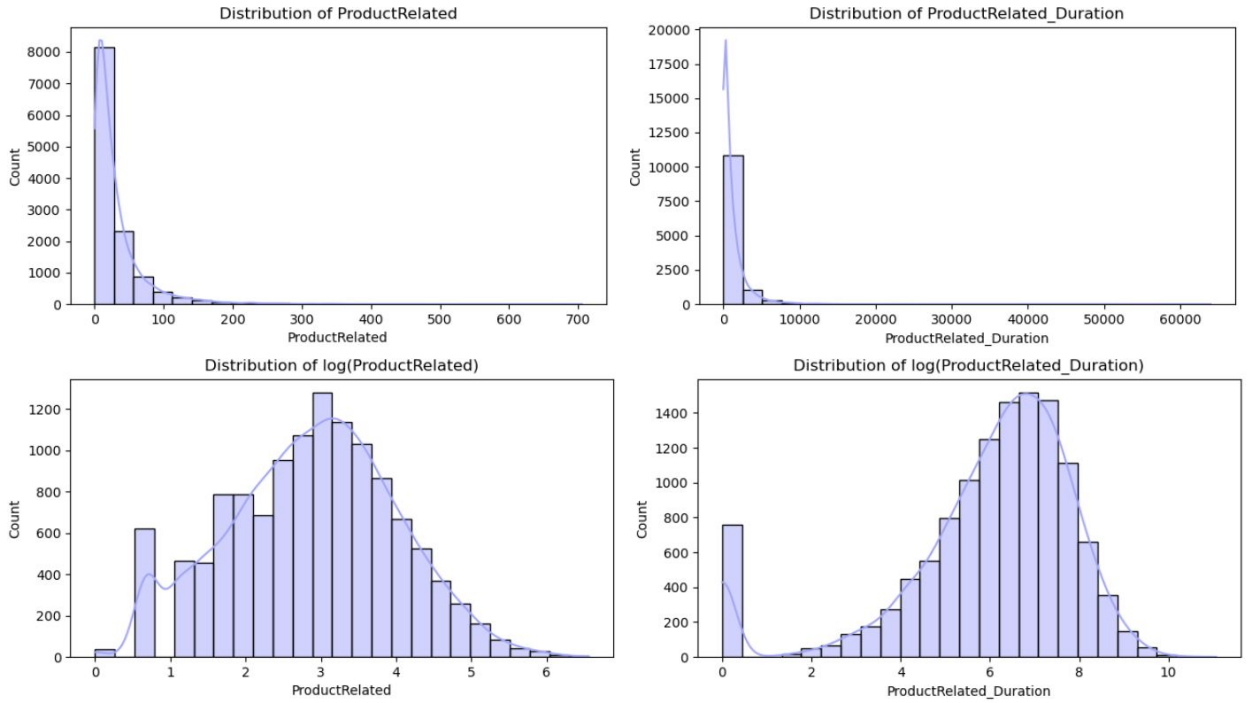


- Specific page and page durations are closely correlated, as expected.
- Bounce Rates & Exit Rates are highly correlated.



- Total Duration: no corr with any except the expected pages.
- Leaving Rate: No correlation except the expected

Distribution Corrections



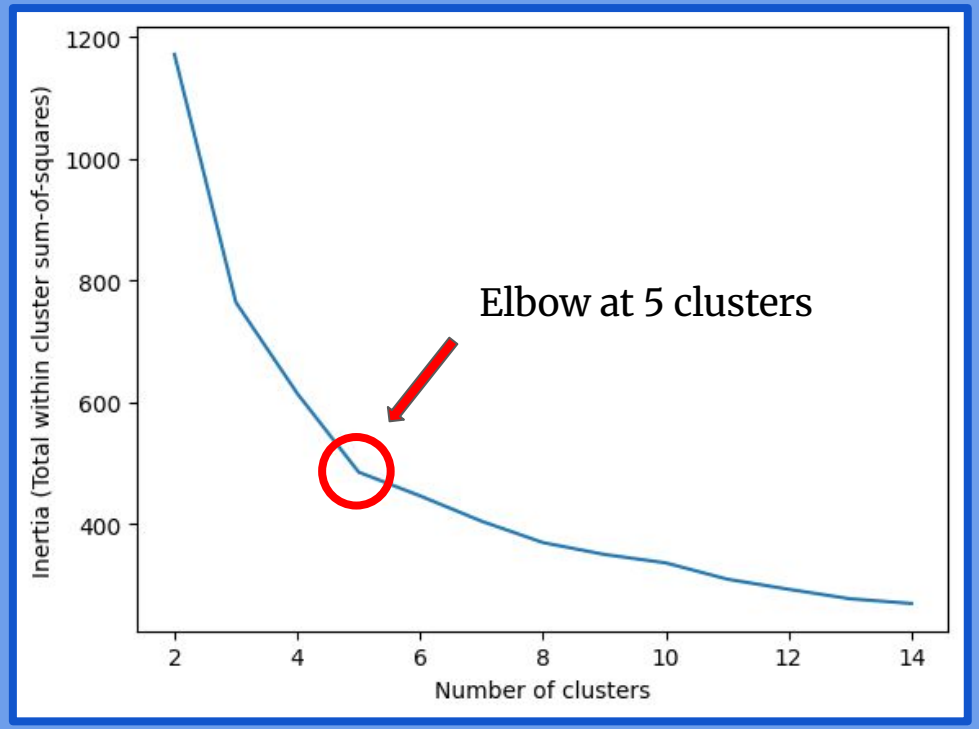
All distributions were clearly skewed, so we performed a log transformation and the only ProductRelated and ProductRelated_Duration showed improvement.

Structuring the data through clustering

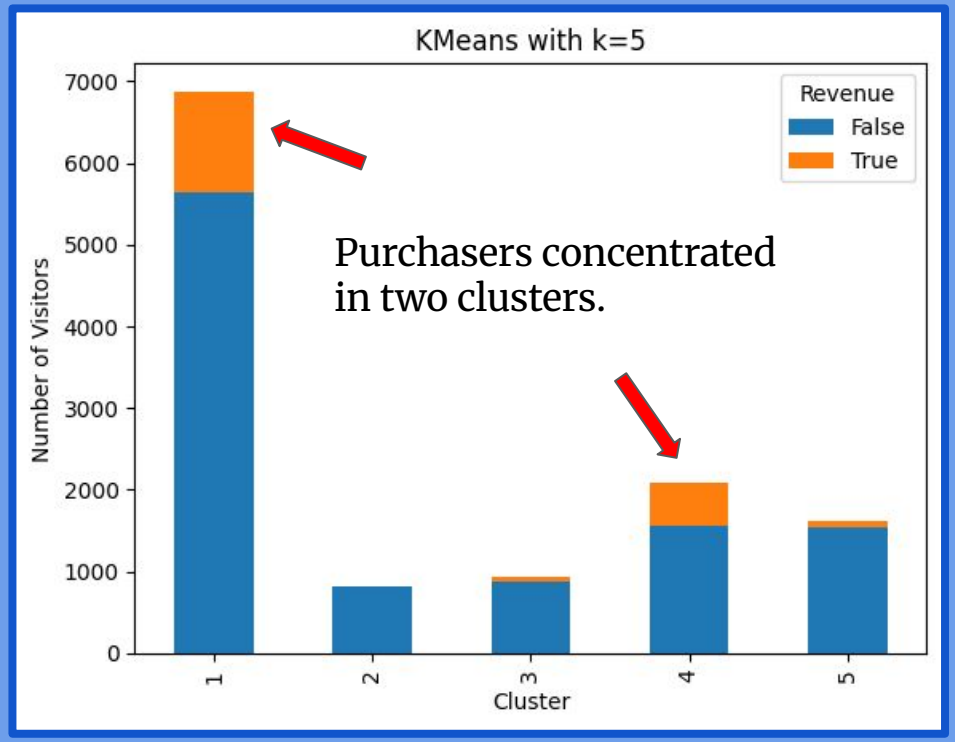
K-Means clustering was applied to the numerical features to see if some natural structure in the dataset could be applied to the classification problem.

First, the numerical features were standardized using min-max scaling.

By the elbow method, we found 5 to be the optimal number of clusters.



Structuring the data through clustering



We found that for the complete dataset, the five clusters did a good job of separating out some non-purchasers from purchasers.

Hence we decided to use K-Means cluster labels as a feature during later modelling.

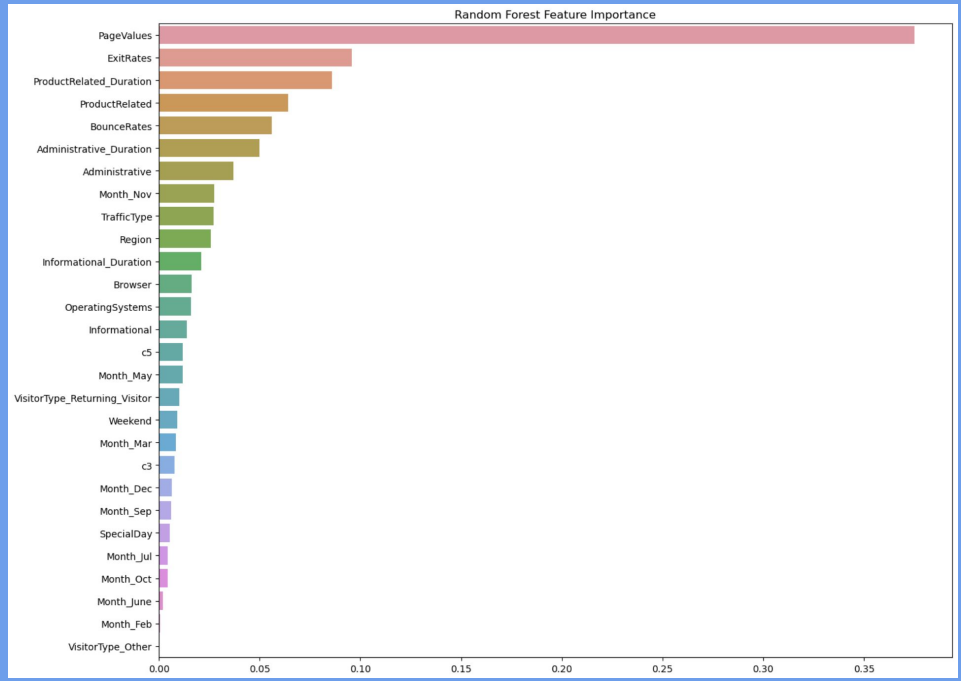
Feature Selection

- Cleaning: Removed duplicates and converted booleans to numeric, one-hot encoding to ‘Month’ and ‘VisitorType’
- Created engineered features - c3 & c5 from clustering.

- Anova F-Test (Numerical)
  - ‘PageValues - F-score of 3895’
  - ‘ExitRates - F-score of 531’
- Chi-Square Test (Categorical)
  - Seasonality: Month_Nov was #1 Categorical predictor (219)
  - c5 scored 72, significantly outperformed native features like Region and TrafficType.

Random Forest & Conclusion

- Random Forest
  - ‘PageValues’ chosen as the best predictor.
  - Followed by ‘ExitRates’ & and ‘ProductRelated_Duration’
- Takeaways
  - ‘PageValues’ ranked as the best predictor and is consistent in all three tests.
  - Churn signals like ‘ExitRates’ show it's just as important as site engagement.

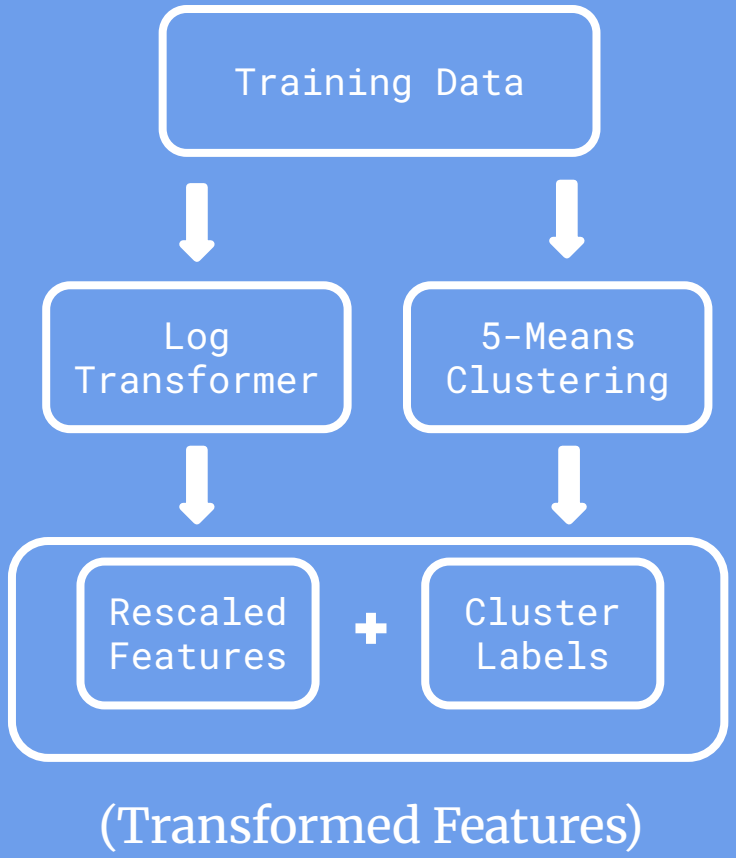


Modeling

Preprocessing Pipeline

As per the results found during EDA, each set of predictors was augmented with a categorical feature that encoded each observation's label using 5-Means clustering.

Additionally, the product-related page predictors were subjected to a log(1 + x) transformation.



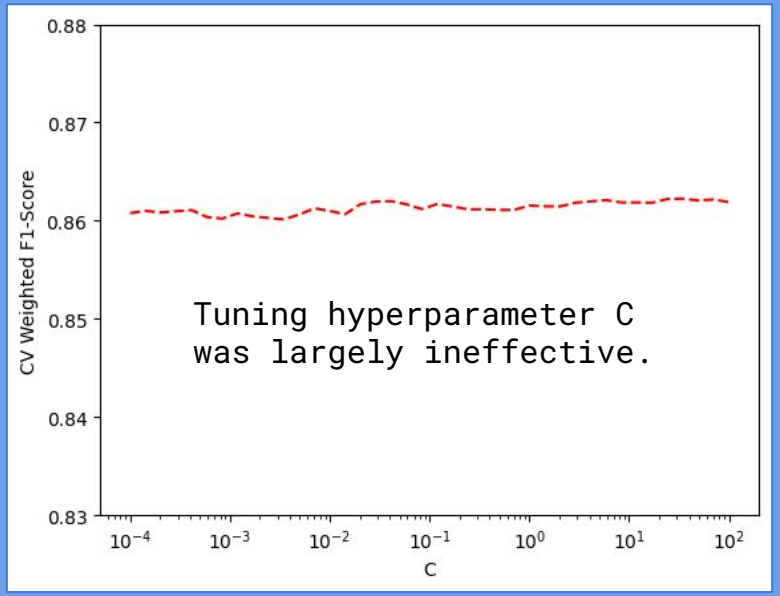
Logistic Regression

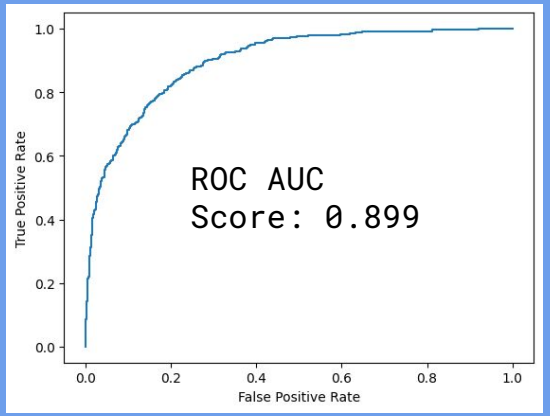
Hyperparameters trained using 10-fold stratified cross-validation:

- C (controlling the strength of L2 regularization on parameter estimates)

Neither tuning the hyperparameter C nor training the model on a resampled dataset appeared to have much effect on the predictions made.

		Predicted	
		2025	34
Actual		237	145





SVM

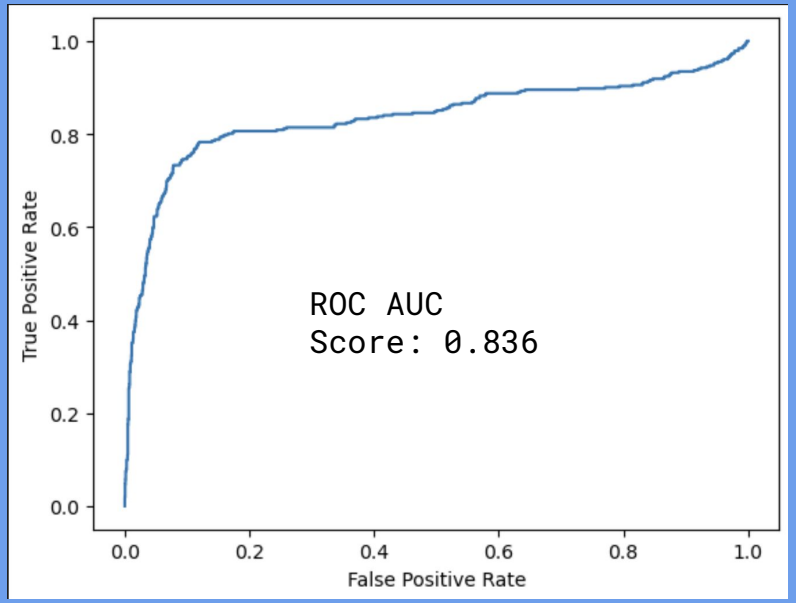
Hyperparameters Trained:

- **C** (Regularization parameter measuring trade-off in decision boundary)
- Kernel (Map data into higher dimensional space, “Linear” vs “RBF” was used.
- **G**amma (Kernel coefficient for RBF - determines how far the strength of training reaches.)

After running GridSearchCV with 10-fold StratifiedKFold, we obtained the following optimal hyperparameters:

- C = 50, Kernel = “RBF”, and Gamma = Scale

Optimal Model Weighted F1 Score: **0.883** out of 1.



	Predicted	
Actual	1995	64
	197	185

K-Nearest Neighbors (KNN)

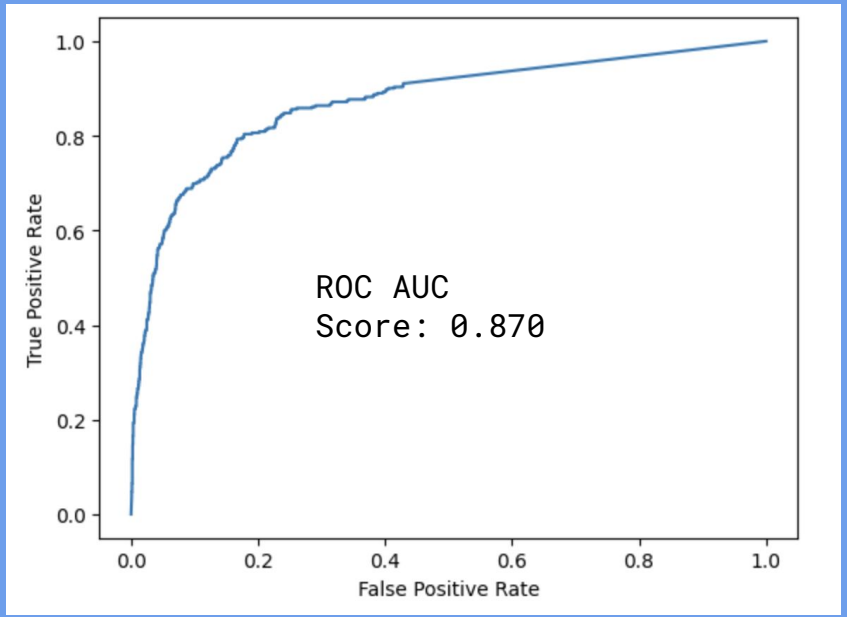
Hyperparameters Trained:

- **P** (If P = 1, Manhattan distance is used, otherwise P = 2 Euclidean distance is used)
- **K** (Testing odd positive integers from [3,25])
- **Weights** (Weight function used when predicting class)

After running GridSearchCV with 10-fold StratifiedKFold, we obtained the following optimal hyperparameters:

- P = 2, K = 13, and Weights = “Distance”

Optimal Model Weighted F1 Score: **0.88** out of 1.



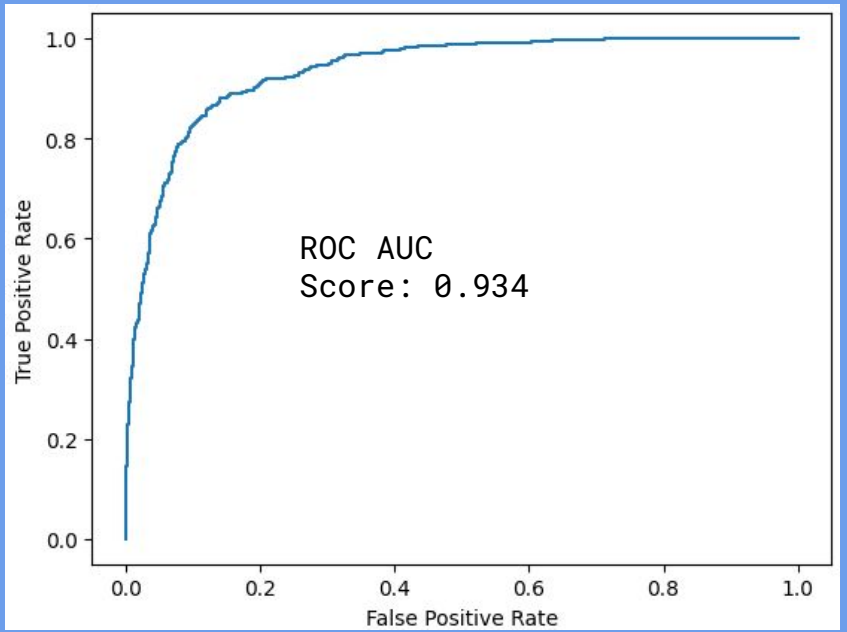
		Predicted
Actual	1993	66
	201	181

Random Forest Classification

Hyperparameters Trained:

- **Estimators** : 500 trees (highest end)
- **Max Depth**: None
- **Max Features:** Sqrt of total features
- **Weights**: Balanced
- **Min Sample Split**: 5 samples before split
- **Criterion**: Entropy over Gini

Optimal Model Weighted F1 Score: **0.906** out of 1.



		Predicted
Actual	1935	124
	115	267

Summary

Logistic Regression:

- Weighted F1: 0.87
- ROC AUC: 0.898
- Time: ~10 seconds

KNN:

- Weighted F1: 0.88
- ROC AUC: 0.87
- Time: ~2 minutes

SVM:

- Weighted F1: 0.883
- ROC AUC: 0.836
- Time: ~23 minutes

Random Forest Classification:

- Weighted F1: 0.905
- ROC AUC: 0.936
- Time: ~80 minutes

Conclusion

- All models had strong prediction capabilities for if a session would result in a ‘No Purchase’.
  - Logistic Regression preferred for low run-time.

- Models consistently struggled with false negatives.
  - High precision but low recall of purchasers.
  - Random Forest was best at balancing precision and recall.
  - Probably owing to imbalance in the dataset.

Applications

- The company could modify their targeted advertising based on features being hit.
  - If Product Duration is high before the Exit Rate
    - Targeted email with a modest coupon
  - Low Page Values & Exit Rate could indicate a sale might not be achieved
    - Company could test high email coupon after Churn
    - Trigger Chat help or UX tweaks
  - Holiday focused, shipping discounts in November
  - Cluster labels to overall adapt targeted intervention