GROUP 7:
Shirley Liu, Haden Loveridge, Aryan Pai,
Twinkle Panda, Kush Patel, Anjali Pillai

EXPLORING TOURISM REPORT

DESCRIPTION

This project focuses on analyzing a dataset, based on user behavior within the travel and tourism industry, specifically focusing on predicting if a user purchases travel products, based on their digital interactions and demographic characteristics. This dataset contains different features like average views on travel related pages, ratings that reflect the user's engagement, the type of location that each user prefers, the number of members in the user's family, and much more. Our main goal is to determine if we can predict the variable, 'Taken_Product', by taking the other features into account, and using different predictive models. The models that were used for this are K-Nearest Neighbors (KNN), Naïve Bayes, Neural Network, Boosting, and XGBoost. These machine learning methods were developed, trained and tested with the dataset, to see which of these models can effectively and accurately predict the target variable and understand user behavior.

The importance of this analysis is to identify the significant impact of the travel and tourism industry by helping companies to understand and predict customer behavior better. Our predictions can allow businesses to create more targeted marketing strategies, improve digital customer behavior and eventually increase revenue. In addition, insights that are gained from this analysis, can be interesting to a broader audience, as it can also be applied to different sectors, like marketers and business analysts across different fields, as the idea of understanding user behavior through online interactions and demographic data is collectively and universally valuable. This approach can help organizations across different industries refine their strategies, enhance customer experiences, and drive business growth.

EXPLORATORY ANALYSIS

DISTRIBUTION OF RESPONSE VARIABLE
The distribution of whether a user makes a purchase is highly skewed, which suggests potential challenges in converting interest into sales.

DISTRIBUTION OF PREDICTOR VARIABLES

- View Frequency Analysis: There is a negative correlation between Yearly_avg_view_on_travel_page and Taken_product. Interestingly, higher engagement with travel pages does not directly lead to increased conversion rates.
- Location Analysis: There is a significant discrepancy between user interest in certain locations and their likelihood of making a purchase. Although there is high interest in locations such as beaches, financial districts, and historical sites, the conversion rates for these categories remain low.
- Follower Analysis: The distribution of website followers reveals that a significant majority of users (64.36%) do not follow the company page, and within this group, only 10.56% made a purchase. On the other hand, among the users who do follow the company page, 30.66% made a purchase. This suggests that following the company page is associated with higher likelihood of making a purchase.
- Family Size Analysis: Although three-member families make up the largest segment, single-member families are more likely to make a purchase. The impact on overall sales could be substantially improved if conversion rate is increased in three and four member families.

MODELS AND INSIGHTS

Features used:

- Dropped UserID since it was distinct across all records and wouldn't contribute to modeling
- Cleaned and typecasted a few columns into boolean columns: working_flag, Taken_product, following_company_page
- For the categorical column preferred_location_type, we used get_dummies to create one-hots to make the column interpretable for our classifiers
- We decided to use all available columns for our analysis post cleaning and were able to reach upto ~99% accuracy

CLUSTER ANALYSIS

In order to dive further into exploratory analysis, k-means cluster analysis was performed. As all the variables are not in the same range, it is necessary to transform and scale the data. There are more than two features in the datasets, it is very hard to perform and visualize cluster analysis on them so we have to reduce the number to two using Principal Component Analysis (PCA).

In order to determine the number of clusters, an elbow graph was used. We chose two clusters to see if customers who buy the ticket next month (target variable) are grouped together and customers who are not are grouped together. The result of this was two visual clusters 0 and 1. After further analysis it was determined that customers in cluster 1, on average give more likes and comments on the travel website and visit the travel page more

often than the customers in cluster 0. Based on the customer behavior, ones in cluster 1 have possibly higher probability of buying the ticket next month.

NAIVE BAYES

The Naïve Bayes model, with all available features, while achieving a reasonable accuracy of 85.1%, exhibited significant weaknesses in identifying users who would engage with the product. The model's precision for the positive class (engagers) was relatively low at 65%, and its recall was especially concerning at just 17%. Since this model often fails to correctly identify users who would engage, we decided to stop exploring this model.

KNN CLASSIFICATION

The KNN model demonstrated comparatively superior performance, achieving an accuracy of 96.3% on the test set. This high level of accuracy indicates that the KNN model is great in distinguishing between users who are likely to engage with the product and those who are not. The precision and recall metrics further underscore the model's reliability, particularly in identifying non-engagers, with a precision of 96% and a recall of 99%. However, while the model is also quite effective at predicting engagers—with a precision of 93%—its recall for this group is slightly lower at 81%, suggesting some room for improvement in reducing false negatives. To get the best results, we tried using a subset of just the top 5 most important variables, but since we saw a consequent decline in accuracy, we decided to keep all the variables in our model. (Figure 1)

Boosting - GBM & XGBoost

The classification task aimed to predict whether a product was taken based on user behavior and preferences.

Model and Results: We utilized two boosting techniques - GBM & XGBoost.

- GBM: Achieved an accuracy of 0.91, precision of 0.93, recall of 0.50, and ROC AUC of 0.70
- XGBoost: Slightly lower accuracy at 0.90, precision at 0.92, recall at 0.41, and ROC AUC of 0.70
- GridSearchCV with XGBoost: Optimized parameters resulted in perfect scores (accuracy, precision, recall, F1, and ROC AUC all at 1.00) for both training and test datasets, indicating excellent model performance

Insights and Key Findings:

(Figure 2) The most important features, as indicated by their F-scores:

- Total likes on outstation check-ins given and received
- Yearly average views and comments on travel pages
- Traveling network rating and check-in frequency

The high accuracy on both training and test sets suggests that the model generalizes well and has effectively captured the patterns in the data. Engagement metrics, such as likes and views, were crucial in predicting the target variable, indicating their strong correlation with product interest. General engagement metrics are more predictive than specific preferences.

Neural Network Implementation and Insights

Approach and Features: To predict whether users would make a purchase on a travel website, we employed a residual neural network (ResNet) due to its effectiveness in handling complex patterns. We used all available features, focusing on those most likely to influence purchasing behavior, such as following_company_page, Adult_flag, and travelling_network_rating.

Model and Results: Our ResNet model, enhanced with dropout layers and an adaptive learning rate, effectively addressed the class imbalance by using weighted binary cross-entropy loss. The model achieved an impressive accuracy of 99.17%, with high precision (98.59%) and specificity (99.73%), indicating strong performance in both predicting purchases and correctly identifying non-purchases.

Insights and Key Findings: Permutation importance and first-layer weight analysis revealed that engagement metrics like following_company_page and demographic indicators such as Adult_flag were the most critical features. Surprisingly, location preferences, expected to be highly influential, had less impact on predictions. This suggests that user engagement and demographics play a more significant role in purchase decisions than specific location preferences.Overall, the neural network provided both high accuracy and valuable insights into user behavior, highlighting areas where travel websites can focus to enhance customer engagement and drive sales.

CONCLUSION:

All the available features were utilized from the dataset, as excluding any of the variables decreased accuracy. Among all the classifier models (Table 1), the champion model was XGBoost, which achieved perfect scores after optimization. Features like, likes and views, were the most important factors, while surprisingly, location preferences had a lesser impact.

REFERENCES:

Table 1: Models Evaluated:

Model Used	Parameters	Accuracy	Precision	Recall	ROC AUC
Naive Bayes	Default params	0.851	False = 86% True = 65%	False = 98% True = 17%	0.74
KNN	Default params	0.963	False = 96% True = 93%	False = 99% True = 81%	0.99
GBM	N_estimators: 300 learning_rate: 0.05 Random_state: 1	0.91	0.93	0.50	0.70
XGBoost	Colsample_bytree: 0.8 learning_rate: 0.1 max_depth: 10 n_estimators: 1000	1.00	1.00	1.00	1.00
Neural Nets		.9917	.9859	.9626	

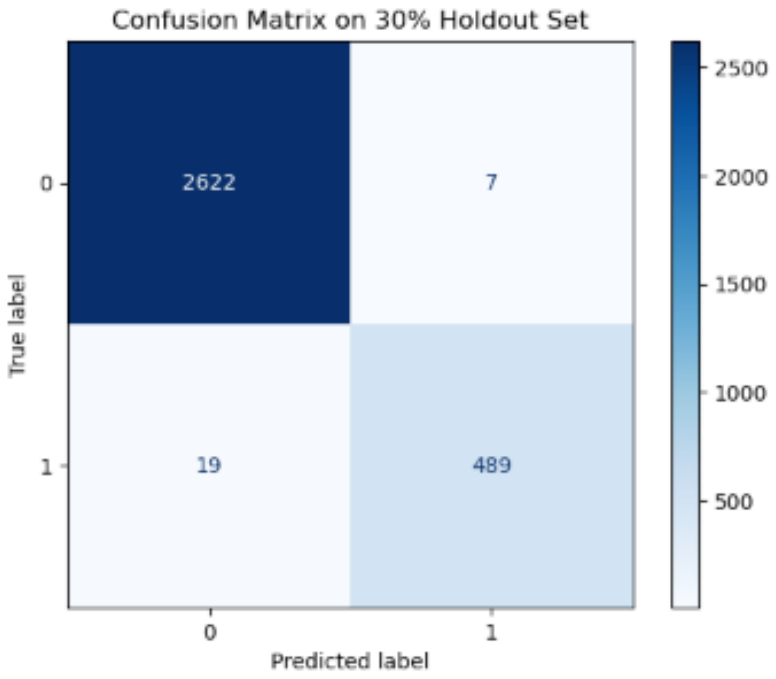

Figure 1: Confusion Matrix for KNN Model

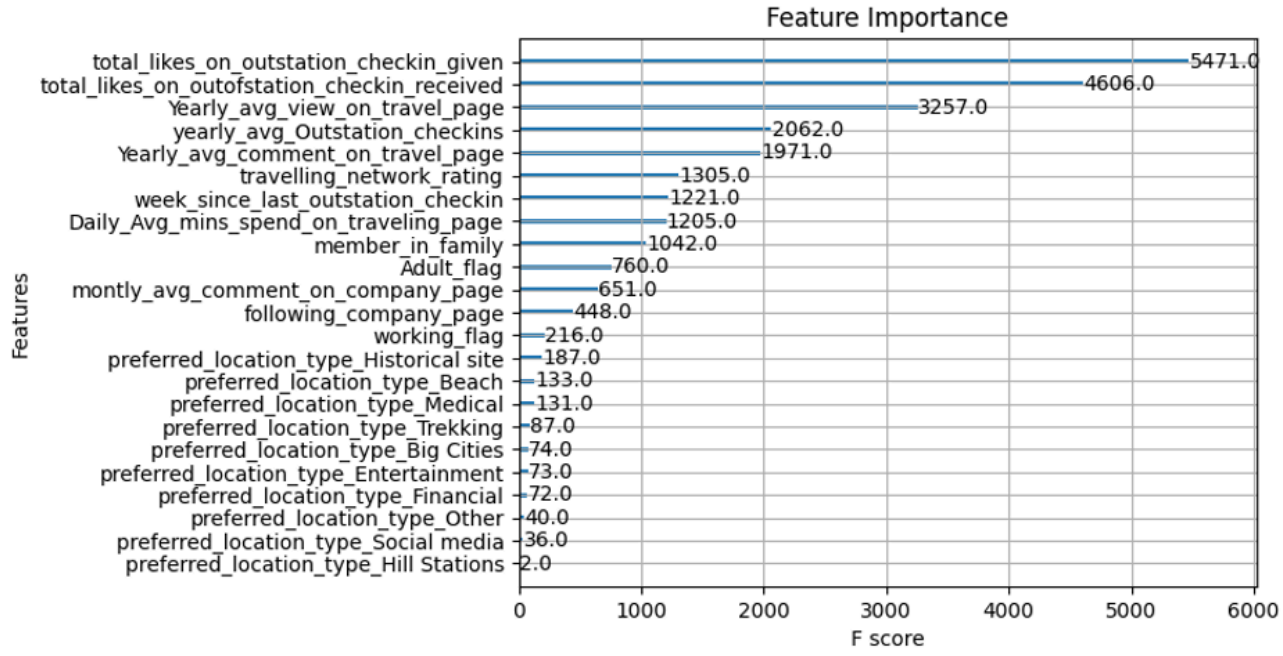

Figure 2: Feature Importance for XG Boost

CITATIONS:

Tourism page engagement. (2023, November 13). Kaggle. https://www.kaggle.com/datasets/ddosad/customer-behaviour-tourism-portal/data

W3Schools.com. (n.d.). https://www.w3schools.com/python/python_ml_knn.asp

Mahadevan, M. (2024, July 10). *Step-by-Step Exploratory Data Analysis (EDA) using Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/