iNeuBytes

      Task-3

  Predictive Analysis and Customer Segmentation for an E- Commerce Platform

**Project Detail**

Can we predict the annual spending of a customer on an e-commerce platform based on their behaviour and demographics? How can these customers segment for effective marketing and personalised user experience.

**Tools that we using in this project :-**

1. Python
2. Power BI/Tableau
3. Scikit-learn
4. Pandas
5. Numpy
6. Matplotlib/Se

**Project Overview**

This project is about customer analytics in the Industry. The project consists of Customer Analytics and Purchased Analytics. It is motivated by Customer Analytics.

* Customer Analytics: The first part of analysis focuses on how to perform customer segmentation. It involves the application of hierarchical and flat clustering techniques for dividing customers into groups. It also features applying the Principal Components Analysis (PCA) to reduce the dimensionality of the problem, as well as combining PCA and K-means for customer segmentation.
* Segmentation: It turns out that it is the most appropriate group of all data points into four segments. Each segment has different characters : 'Well-Off', 'Fewer-Opportunities', 'Standard', 'Career\_Focused'. "Fewer-Opportunities" segment is the largest. Almost 40 percent of customers belong to this segment.

                 Task-3 Questions & Answers

These questions were related to Major Project, therefore Answer all the question are :

**Q.1 - What is the business model of the e-commerce platform you're working with?**

Ans - The business model of the e-commerce platform can be based on selling products or services online to customers. The platform may generate revenue through various streams such as:

1. Sales Revenue:

The platform earns revenue by selling products or services directly to customers. This can be through a marketplace model where the platform facilitates transactions between buyers and sellers, or through a direct-to-consumer model where the platform owns and sells its own products.

1. Commission or Transaction Fees:

If the platform operates as a marketplace, it may charge a commission or transaction fees from sellers for each successful sale made on the platform. This can be a percentage of the transaction value or a fixed fee per transaction.

1. Subscription or Membership Fees:

The platform may offer premium features or services to customers and charge a recurring subscription or membership fee to access those enhanced features or benefits.

1. Advertising Revenue:

The platform may generate revenue by displaying advertisements to users. This can include sponsored product listings, banner ads, or targeted ads based on user behaviour and preferences.

1. Data Monetization:

The platform may collect and analyse customer data to derive insights and trends. This data can be anonymized and sold to third parties for market research, targeted advertising, or other data-driven purposes.

1. Fulfilment Services:

If the platform provides fulfilment services, such as warehousing, packaging, and shipping, it may charge fees for these services to sellers who utilise them.

**Q.2 - What kind of data preprocessing and cleaning was required for the Online Retail II Dataset?**

Ans - The Online Retail II Dataset may require several data preprocessing and cleaning steps to handle missing, erroneous, inconsistent data, and outliers. Here are some common data preprocessing and cleaning tasks that can be performed on the dataset:-

1. Handling Missing Values: Check for missing values in the dataset and decide how to handle them. You can either remove the rows or columns with missing values or impute the missing values using techniques like mean, median, or regression imputation.
2. Handling Erroneous or Inconsistent Data: Identify any erroneous or inconsistent data entries that do not align with the expected data patterns or rules. For example, you might check for negative quantities or prices that are not possible in the context of the dataset. These entries can be corrected, removed, or flagged for further investigation.
3. Data Type Conversion: Convert data types to their appropriate formats. For example, dates should be converted to datetime objects, categorical variables should be encoded appropriately (e.g., one-hot encoding or label encoding), and numeric variables should be converted to numeric data types.
4. Removing Duplicates: Check for duplicate records in the dataset and remove them to ensure the integrity of the data. Duplicates can occur due to data entry errors or system glitches.
5. Outlier Detection and Treatment: Identify outliers in numerical variables that deviate significantly from the expected range. Outliers can be detected using statistical methods such as z-score, IQR (Interquartile Range), or visualisation techniques. Depending on the nature of the outliers and the context of the data, they can be removed, winsorized (capped at a certain threshold), or treated through transformations like log transformation.
6. Feature Scaling or Normalisation: Scale numeric features to a similar range to avoid dominance of certain features during modelling. Common scaling techniques include min-max scaling or standardisation (z-score normalisation).
7. Handling Skewed Data: Address skewness in numeric variables by applying transformations like log transformation, square root transformation, or Box-Cox transformation. Skewed data can impact the performance of certain models.
8. Data Aggregation: If necessary, aggregate the dataset at a higher level (e.g., customer level, monthly level) by summarising transactional data. This can be done by grouping the data based on specific attributes and calculating aggregations like sum, mean, or count.

**Q.3 - How did you visualize and interpret the data distributions and relationships using Power Bl/Tableau?**

Ans - I can provide you with an overview of how you can visualize and interpret data distributions and relationships using Power BI or Tableau:

* 1. Data Distribution Visualization:
  + Histogram: Use histograms to visualize the distribution of a numeric variable. Power BI and Tableau provide built-in histogram charts that group data into bins and display the frequency or count of data points within each bin.
  + Box Plot: Create box plots to visualize the distribution of a numeric variable across different categories or groups. Box plots show the median, quartiles, and potential outliers, providing insights into the spread and skewness of the data.
  + Bar Chart: Use bar charts to display the distribution of a categorical variable. You can aggregate the data and show the count or percentage of each category.
* 2. Relationship Visualization:
  + Scatter Plot: Create scatter plots to visualize the relationship between two numeric variables. Power BI and Tableau allow you to plot data points on a Cartesian coordinate system, where each point represents the values of two variables.
  + Line Chart: Use line charts to explore the relationship between two variables over time. Line charts show trends and patterns in the data by connecting data points with lines.
  + Heatmap: Create a heatmap to visualize the relationship between two categorical variables. Heatmaps use color intensity to represent the strength or frequency of the relationship between different categories.
* 3. Interactive Exploration:
  + Power BI and Tableau provide interactive features like filtering, drill-down, and highlighting to enable deeper exploration of data distributions and relationships. Users can interact with the visualizations to focus on specific subsets of data or drill down to more granular levels

**Q.4 - What new features did you engineer from the existing dataset and why?**

Ans - The potential feature engineering techniques that you can consider for the Online Retail II Dataset or any e-commerce dataset:

1. Monetary Value:

Calculate the monetary value of each transaction by multiplying the quantity of items purchased by the price. This new feature can capture the spending behavior of customers.

1. Recency:

Calculate the number of days between the most recent purchase and a reference date. This feature can capture how recently a customer has interacted with the platform and may indicate their level of engagement.

1. Frequency:

Calculate the total number of transactions made by each customer. This feature can represent the level of activity or loyalty of customers.

1. Average Order Value:

Calculate the average value of each customer's transactions. This feature can capture the typical spending behavior of customers.

1. Time-based Features:

Extract additional time-related features from the transaction timestamps, such as hour of the day, day of the week, month, or season. These features can capture temporal patterns and help identify any time-dependent behaviors or trends.

1. Categorical Features:

If the dataset includes categorical variables like product categories or customer segments, you can create binary or count-encoded features for each category. These features can capture the preferences or characteristics of customers or products.

These are just a few examples, and the specific feature engineering techniques will depend on the dataset and the insights you are trying to capture. It's important to understand the data, explore the relationships between variables, and use domain knowledge to guide the creation of meaningful and relevant features.

**Q.5 - Which regression models did you test for predicting the annual spending of a customer?**

Ans - When predicting the annual spending of a customer on an e-commerce platform, several regression models can be tested to find the best performing one. Here are some commonly used regression models that you can consider:

1. Linear Regression:

This is a basic regression model that assumes a linear relationship between the input variables and the target variable. It can provide insights into the direction and magnitude of the relationships between the features and the annual spending.

1. Decision Tree Regression:

Decision tree regression models create a tree-like structure to make predictions. They can capture non-linear relationships and interactions between features. However, they may be prone to overfitting.

1. Random Forest Regression:

Random forest regression combines multiple decision trees to make predictions. It helps to reduce overfitting and improves prediction accuracy by averaging the results of individual trees.

1. Gradient Boosting Regression:

Gradient boosting regression is an ensemble method that combines weak learners (usually decision trees) in a sequential manner. It iteratively corrects the errors made by previous models to improve overall prediction accuracy.

1. Support Vector Regression:

Support Vector Regression (SVR) is a regression model that uses support vector machines to find the best hyperplane that fits the data. It can handle non-linear relationships through the use of kernel functions.

**Q.6 - What metrics did you use to evaluate the performance of the predictive models?**

Ans - To evaluate the performance of predictive models for predicting the annual spending of a customer on an e-commerce platform, several metrics can be used. Here are some commonly used evaluation metrics for regression models:

1. Mean Absolute Error (MAE):

MAE measures the average absolute difference between the predicted values and the true values. It provides a straightforward interpretation of the average prediction error and is less sensitive to outliers.

1. Mean Squared Error (MSE):

MSE measures the average of the squared differences between the predicted values and the true values. It penalizes larger errors more than MAE and is commonly used in regression models.

1. Root Mean Squared Error (RMSE):

RMSE is the square root of MSE. It provides an interpretable metric in the same unit as the target variable, making it easier to compare across different models.

1. R-squared (R2) Score:

R-squared measures the proportion of the variance in the target variable that is explained by the regression model. It ranges from 0 to 1, with higher values indicating better fit. However, R-squared may not be sufficient for complex models or when there are non-linear relationships.

**Q.7 - How did you apply clustering techniques for customer segmentation? What were the results?**

Ans - To apply clustering techniques for customer segmentation, you can use algorithms like K-means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), or hierarchical clustering. Here's a general approach to applying clustering for customer segmentation:

1. Data Preparation:

Prepare the dataset by selecting relevant features for segmentation. These features can include customer demographics, purchase behavior, or any other attributes that can differentiate customers.

1. Feature Scaling:

Normalize or standardize the features to ensure that they are on a similar scale. This step is important because clustering algorithms are sensitive to the scale of the variables.

1. Clustering Algorithm Selection:

Choose an appropriate clustering algorithm based on the nature of the dataset and the goals of the segmentation. K-means is a popular choice for its simplicity, while DBSCAN is useful for identifying clusters of varying shapes and sizes.

1. Determine the Optimal Number of Clusters:

If using K-means, you can utilize techniques like the elbow method or silhouette analysis to determine the optimal number of clusters. This step helps to find a balance between the complexity of the segmentation and the interpretability of the clusters.

1. Run the Clustering Algorithm:

Apply the chosen clustering algorithm to the prepared dataset. Each customer will be assigned to a specific cluster based on the algorithm's clustering rules and similarity metrics.

The specific results of customer segmentation using clustering techniques will depend on the dataset, chosen algorithm, and the underlying patterns in the data. The goal is to identify distinct customer segments with similar characteristics, behaviors, or preferences. These segments can then be used for targeted marketing strategies, personalized user experiences, or other business decisions.

It's important to note that customer segmentation is an iterative process, and the interpretation and analysis of the clusters should be done in collaboration with domain experts or stakeholders. The insights gained from the segmentation can help in tailoring marketing campaigns, optimizing customer experiences, and making data-driven business decisions.

**Q.8 - How would you interpret the results obtained from the model in a business context?**

Ans - Interpreting the results obtained from the predictive model in a business context involves translating the technical findings into meaningful insights and actionable recommendations that can drive business decisions. Here are some steps to effectively interpret the results:

1. Understand the Model's Performance:

Begin by explaining the performance metrics used to evaluate the model, such as MAE, MSE, RMSE, or R-squared. Clearly state how well the model predicts the annual spending of customers.

1. Identify Important Features:

Discuss the importance of different features in the predictive model. Highlight which features have the most significant impact on annual spending. This information can help the business understand the key drivers of customer spending.

1. Customer Segmentation Insights:

If clustering techniques were applied, explain the identified customer segments and their characteristics. Provide insights into the distinct groups of customers and their spending behaviors. For example, high-value customers, low-value customers, frequent shoppers, etc.

1. Business Implications:

Translate the model's findings into actionable business implications. For example, which customer segments are the most valuable and should be prioritized for targeted marketing efforts? Are there any customer segments that need special attention to improve their engagement and spending?

1. Personalization Opportunities:

Explore opportunities for personalized user experiences based on the predictive model's insights. For instance, the model may suggest tailoring product recommendations or offers to different customer segments to enhance customer satisfaction and increase sales.

**Q.9 - How can the insights derived from this project be beneficial for the e-commerce platform's business strategy?**

Ans - The insights derived from this project can be highly beneficial for the e-commerce platform's business strategy in several ways:

1. Customer Segmentation:

By identifying distinct customer segments based on spending behavior and other characteristics, the platform can tailor its marketing strategies and offerings to each segment. This enables personalized communication, targeted promotions, and enhanced customer experiences, leading to increased customer satisfaction and loyalty.

1. Targeted Marketing Campaigns:

The predictive model can help the platform identify high-value customer segments that contribute significantly to the revenue. With this knowledge, the platform can develop targeted marketing campaigns to attract and retain these valuable customers, maximizing their lifetime value.

1. Product Recommendations:

By understanding customer preferences and purchase patterns, the platform can improve its product recommendation engine. Utilizing the predictive model's insights, personalized recommendations can be provided to customers, enhancing their shopping experience and increasing the chances of cross-selling and upselling.

1. Inventory Management:

The predictive model's revenue forecasting capability can assist in optimizing inventory management. By forecasting future sales based on customer behavior, the platform can ensure sufficient stock levels for high-demand products and reduce inventory costs for low-demand items, leading to improved operational efficiency.

1. Pricing Strategies:

Insights from the predictive model can guide pricing strategies for different customer segments. The platform can identify price-sensitive segments and offer targeted discounts or promotions to stimulate their purchasing behavior.

**Q.10 - What did you learn about the data science project lifecycle throughout this project?**

Ans - Throughout this project, several key aspects of the data science project lifecycle were encountered and learned. Here are some important takeaways:

1. Business Understanding:

The project began with a focus on understanding the business model of the e-commerce platform, its revenue streams, customer base, and the importance of customer segmentation and spending prediction. This step highlighted the significance of aligning data science goals with the business objectives.

1. Data Collection and Cleaning:

The process of data collection involved obtaining the Online Retail II Dataset, a popular dataset containing transaction data for a UK-based online retail. Cleaning the dataset was crucial to handle missing, erroneous, and inconsistent data, as well as identifying and treating outliers. This step emphasized the importance of data quality and preprocessing for accurate analysis.

1. Exploratory Data Analysis:

Power BI/Tableau was used to perform exploratory data analysis, uncovering insights and patterns within the dataset. Visualizations and data summaries were created to understand data distributions, relationships among variables, and potential trends or anomalies.

1. Feature Engineering and Transformation:

New features were engineered from the existing dataset to capture additional insights that might be relevant for the predictive model. This involved creating meaningful variables or transforming existing ones to better represent underlying data patterns. Feature engineering highlighted the importance of domain knowledge and creativity in enhancing model performance.

1. Predictive Model Building:

Various regression models such as linear regression, decision tree regression, random forest regression, etc., were tested for predicting the annual spending of customers. Model building involved training, tuning, and evaluating these models to select the best performing one. This phase emphasized the iterative nature of model development and the need for proper evaluation and selection.

1. Model Evaluation and Customer Segmentation:

The models were evaluated using metrics like MAE, MSE, RMSE, and R-squared to assess their performance. Additionally, clustering techniques such as K-means, DBSCAN, or hierarchical clustering were applied for customer segmentation. Evaluation metrics and segmentation techniques provided insights into model effectiveness and customer groupings.

1. Interpretation and Communication:

The results obtained from the models and segmentation were interpreted in a business context, deriving insights and drawing conclusions. The findings were communicated effectively through written reports and presentations to stakeholders, emphasizing the importance of clear and concise communication skills.

Throughout the project lifecycle, it became evident that data science projects are iterative and require a combination of technical skills, domain knowledge, and effective communication. The importance of understanding the business context, ensuring data quality, feature engineering, model evaluation, and interpretation of results were key elements in delivering actionable insights and making informed business decisions.

**Task - 3 Completed**