**FRAUD AND RISK ANALYTICS**

**PROJECT**

**Instructions for the submission:**

* Please maintain the following: Font - Times New Roman, Font Size - 12, Line Spacing - 1.5

|  |  |  |  |
| --- | --- | --- | --- |
| S.No. | Assessment | Submission Format | Marks |
| 1 | Univariate Analysis | Text+ Screenshots | 10 marks |
| 2 | Derived Metrics | Text+ Screenshots | 10 marks |
| 3 | Bivariate Analysis | Text+ Screenshots | 10 marks |
| 4 | Model Analysis | Text+ Screenshots | 10 marks |
| **Project Maximum Marks** | | | **40 marks** |

I have uploaded the Python File at Google Drive, link below:

[**https://drive.google.com/file/d/1COF8tn1lX6o9h2KbtiZD9Rjr7eJu5H25/view?usp=sharing**](https://drive.google.com/file/d/1COF8tn1lX6o9h2KbtiZD9Rjr7eJu5H25/view?usp=sharing)

|  |  |
| --- | --- |
| **Question 1** | A key step………. |
| **Marks** | 10 marks |
| **Word Count** | N/A |
| **Your Answer** | *Write your answer here + Screen Shots of code/ Excel*  **Visually analyze the following variables independently and draw actionable insights. (Initial Python Code Below)**  fraudDataSetOrignal = pd.read\_excel('Fraud+Analytics+Dataset.xlsx')  fraudDataSet = fraudDataSetOrignal.copy()  fraudDataSet.info() #Checking the types of values/ types / Number of records  fraudDataSet.head(5) #Checking the column Names and value   * 1. **- Age:**   **Finding the minimum and maximum values of Age for bucketing:**    **Creating the buckets , adding new column for AgeRange and drawing insights from the buckets:**    **Plotting the buckets and finding the percentage of transactions done by those people on a Horizontal Bar:**    **Insight:**  We can clearly see from the stats and plot above that people between **Ages 30-39** does maximum number of transactions which is around **40%** of the total transactions. After this Age group **18-29** does maximum number of Transactions(36.7%).  *----------------------------------------------------------------------------------------*   * 1. **Sex**   First I replace the values M with Male and F with Female then, I created insights for the same and also calculated percentages of transactions done by each category (Male and Female), furthermore I plotted them on **PIE** chart.      **Insight:**  We can clearly see from the stats and pie above that Male are doing more transactions than female. Male have done around **88293(58.4%)** transactions and Females have done around **62819(41.6%)** transactions from a total of from the total 15112 transactions  *----------------------------------------------------------------------------------------*   * 1. **Purchase Value**   For Purchase Value I found out the minimum and maximum first and then created bins taking a range of 10 at a time. Added a column for PurchaseRange and found insights from the Total transactions in that purchase range the percentage of transactions in that range, plotted a **histogram** for finding the insights. Code Below:          **Insight:**  We can clearly see from the numbers and histogram above that maximum number of transactions are done for Purchase Range **21-40** which is  **60766(40.2%)** and next is **41-60** which is **41043(27.1%)**  ----------------------------------------------------------------------------------------   * 1. **Browser**   For Browser I did straight analysis on which browser was used to make transactions based on value count code and also found out the percentage of transactions done on each browser, finally I plotted a **PIE** of percentages and represented the insights below:      **Insight:**  We can clearly see from the above figures and PIE that maximum transactions are done by **Chrome - 61432 (40.7%)** browser followed by  **IE -36727 (24.3%)**, **Safari - 24667 (16.3%), Firefox - 24610 (16.3%)** and least are done on **Opera - 3676 (2.4%)**.  ----------------------------------------------------------------------------------------   * 1. **Source**   For Source also, I did straight analysis on what was the source of making transactions based on value count code and also found out the percentage of transactions through each source, finally I plotted a **horizontal Bar Chart** of percentages and represented the insights below:      **Insight:**  We can clearly see from the above figures and Horizontal Bar above that the source **SEO – 60615(40.1%)** and **Advertisement - 59881**  **(39.6%)** bring the maximum transactions and the least are done via  **Direct – 30616 (20.2 %)** source. |

|  |  |
| --- | --- |
| **Question 2** | Now Before ……… |
| **Marks** | 10 marks |
| **Word Count** | N/A |
| **Your Answer** | *Write your answer here + Screen Shots of code/ Excel*  **2.1 - Hypothesis 1: Transactions that occur within the same date as sign-up have a higher probability of being fraudulent in nature.**  For this first I extracted Date from the SignUpDate timestamp column,  secondly I created a **derived metrics** called as **‘signup\_Equals\_PurchaseDate’**  column which will have Yes value if PurchaseDate is equal to Sign up date and No if not.  Then I tried to find out the number of Yes and No for **signup\_Equals\_PurchaseDate** in the Dataset and also found out the percentages, code and result below:      Furthermore, I tried to find the transactions that are fraudulent in the dataset as well,  codes and insights below:    **Insight:**  We can clearly see from the above figures and Pie diagrams that **142908 (95%)** transactions happened when the **SignUp Date was same as purchased date**, also if we  consider the dataset with all fraudulent transactions we can see that **7630(54%)** happened when the signup date was same as purchase date.  Therefore, **Hypothesis 1**: Transactions that occur within the same date as sign-up have a higher probability of being fraudulent in nature, is **ACCEPTED**.  -------------------------------------------------------------------------------------------------------  **2.2** - **Hypothesis 2**: Transactions that occur after midnight and before 3 am have a higher likelihood of being fraudulent in nature.  For this first I created **derived metric** called as **‘PurchaseBetweenMidnightto3AM’**  column which will have **Yes** value if Purchase Time is between 12:00AM and  3:00AM and **NO** if Not.  Then I tried to find out the number of Yes and No for **‘PurchaseBetweenMidnightto3AM’** in the Dataset and also found out the percentages, code and result below:    Furthermore, I tried to find the transactions that are fraudulent in the dataset as well,  codes and insights below:    **Insight:**  We can clearly see from the above figures and Pie diagrams that **132361**  **(88%)** transactions **did not** happen between 12:00AM and 3:00AM, also if we  consider the dataset with all fraudulent transactions we can see that **12455**  **(88%)** f**raud transactions did not** happen between 12:00AM and 3:00AM  Therefore, **Hypothesis 2**: Transactions that occur after midnight and before 3 am  have a higher likelihood of being fraudulent in nature is **REJECTED.** |

|  |  |
| --- | --- |
| **Question 3** | Create a……….. |
| **Marks** | 10 marks |
| **Word Count** | N/A |
| **Your Answer** | *Write your answer here + Screen Shots of code/ Excel*  **3.1**. Create a correlation matrix for all the numerical variables and analyse the results. Also, check for multicollinearity. (5)  I have considered the numerical variables Purchase Value and Age in this Correlation test and also checked correlation among them and with Fraud Indicator. Codes and Heatmap below:      **Insight:**  I can clearly see from the above correlation and Heatmap that correlation coefficient is **0.002370(Purchase Value and age), 0.001011**  **(PurchaseValue and Fraud)** and **0.006624(Age and Fraud),** all these value are **too**  **less** and we can easily say that there is **Negligible Correlation** between  the variables,  **[Checking for multicollinearity]**  I used the VIF function to check for Multicollinerity, codes and insights below:    **Insight:**  I can clearly see from the above VIF values that all are below 5 therefore we can infer that the values are **not** Multicolliner.  ------------------------------------------------------------------------------------------------  **3.2 Do pairwise analysis of every categorical variable with the outcome variable and draw actionable insights.**  There are three categorical variables –, **Sex, Browser and Source**, below is the analysis between these variables(pairwise) and actionable insights.   1. **(Sex, Source) with respect to Fraud**:     **Insight:**  We can clearly see from Association and Pivot table above that  Fraudulent transactions are more done by **Male (8434)** from which  **Ads (3227)** and **SEO (3214)** are the sources for higher fraudulent  transactions and **Direct (1993)**.  **Females (5717)** did lesser fraud transactions but the weightage for Ads,  SEO remains high and less for Direct.  Company will have to make rule and check transactions done by **Males**  **coming from Ads and SEO**. Next Priority is **Females coming from**  **Ads and SEO.**  ----------------------------------------------------------------------------------------   1. **(Browser, Sex) with respect to Fraud**:     **Insight:**  We can clearly see from the Association and Pivot Table above that  Fraudulent Transactions **Males** and **Females** on **Chrome Browser (3642, 2427 respectively)**, followed by **IE** browser then follows **Firefox, Safari and Opera.**  Count of **Males** is more for each browser.  Therefore, we can use Fraud Rule or try to analyze **Male** customers  using **Chrome and IE** browser instantly and then move to Firefox,  Safari and Opera browser customer to detect and eliminate fraud.  ------------------------------------------------------------------------------------------------   1. **(Source, Browser) with respect to Fraud:**         **Insight:**  We can clearly see that Fraudulent transactions are more occurring from **Chrome Browser** – coming from **Source -** **Ads(2411) and SEO(2157) and Direct(2157)**, followed by IE, Firefox and then Safari and Opera. We need to put Fraudulent  rule for people using Chrome and coming through Source Ads and SEO first, followed by Direct source. Then move to IE, Firefox, Safari and Opera in the end. |

|  |  |
| --- | --- |
| **Question 4** | The final……….. |
| **Marks** | 10 marks |
| **Word Count** | N/A |
| **Your Answer** | *Write your answer here + Screen Shots of code/ Excel*  **Calculate the following evaluation metrics for the two confusion matrix given above:**   1. **Accuracy** 2. **Recall** 3. **Precision**   For approaching this part, I wrote code in Python for calculating all the metrics and passed the value as per confusion metrics for KNN and  Decision Tree separately:    Calculation for KNN and Decision Tree Below:    **KNN:**  Accuracy: 0.95  Recall: 0.54  Precision: 0.86  **Decision Tree:**    **Decision Tree:**  Accuracy: 0.90  Recall: 0.57  Precision: 0.48  For the given problem, analyze the model metrics you have calculated so far and select the best model for this use case. Justify your answer.  For the above question I added one more metric, below:    **KNN:**    F1 Score for KNN is 0.66  **Decision Tree:**    F1 Score for Decision Tree is 0.52  **Insight/Analysis:**   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **Accuracy** | **Recall** | **Precision** | **F1-Score** | | **KNN** | 0.95 | 0.54 | 0.86 | 0.66 | | **Decision Tree** | 0.9 | 0.57 | 0.48 | 0.52 |  * Recall is a metric for how many of the Actual Positives our model captured through labeling it as True Positive. * Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. * We mostly use Recall for modelling of Fraudulent Transactions. Here we see that the recall value of Decision Tree is more than KNN, but KNN has better Accuracy and Precision. * **Rather than going for final accuracy on the dataset, we need to care about catching most of the fraud cases (Recall), while keeping the cost at which this is achieved under control (precision). Usually, this is captured in the F1 score.** * The Accuracy and Precision Scores are high but the recall scores are low, to save the costs for fraud transaction we need to take recall into consideration and we cannot rely only on Precision and Accuracy. * **F1 Score** is a better metric to use if we need to create a optimum balance between precision and recall. The F1 score tries to take this into account, giving more weight to false negatives and false positives while not letting large numbers of true negatives influence the score.   **Therefore, we need can say from above that KNN model is better than Decision Trees as KNN has a FI Score of 0.66 and Decision Tree FI Score is 0.52.** |