**AEGIS SCHOOL OF DATA SCIENCE**

CAPSTONE PROJECT: FINAL SUBMISSION

|  |  |
| --- | --- |
| Name | Nikita Dange, Sanjiv Sharan, Shriti Datta |
| Batch | Eg. DP1812 |
| Roll Number |  |
| Other Group Members (if any) |  |
| Project Title | Customer and Product Profiling |
| Project Description | Perform Customer Analytics on e-commerce data to create solutions which help organization to increase sales by spending less money.   * Classifying customers into segments. * Anticipate the purchases that will be made by a new customer. |
| Final deployment of the project | Dashboard depicting customer segmentation |
| List your major learnings from this project | Eg:   * Understood e-commerce domain * Understood different classification techniques * Understood text analytics * Learned different plotting diagrams * Dashboard creation * Cloud implementation |
| Major tools used | Eg. PowerBI, AWS, Python – including Packages- seaborn, Matplotlib, numpy, sklearn, wordcloud, IPython.display |
| Final number of data points used in the project | 532610 records |
| Final number of data points used for training | 426088 |
| Final number of data points used for testing / validation | 106522 |
| Mention cross validation techniques used, if any | Cross-validation k-fold=5 |
| Data sourced from | Online Retail-ecommerce (UK Retailers)- One Year data  UCI Machine Learning Repository |
| Methods used to source the data |  |
| Challenges faced to source data | We used data provided on UCI Machine Learning Repository |
| Describe the efforts spent and challenges overcome during data preparation, and the steps used | * Analyze significance of all data features and remove unwanted features * Sanity check of the important data |
| Outline the major steps used to implement the project | Workflow |
| List the major machine learning techniques used / implemented in your project | SVM, KNN, Logistic regression, Decision tree, Random Forest, Adaboost, Gradient Boosting |
| Mention the major challenges faced and overcome during model creation | 1. Imbalanced datasets 2. Applying different classification techniques 3. Hyperparameter tuning to optimize the model |
| List the techniques used to explore the data | 1. Identify null values 2. Frequency distribution 3. Remove duplicate values 4. Explore data through visualization |
| List the outcome of data exploration | 1. Cancelled orders were eliminated 2. Duplicate values eliminated |
| List any unsupervised learning technique used during data exploration | 1. Creating word Cloud 2. Created 5 Product Clusters 3. Created 10 Customer Clusters |
| How many features were finally identified to create the model | 8 features |
| Describe your Feature Engineering steps, if any | 1. Cancelled ordered identified 2. Product categories |
| List and describe any feature reduction techniques used | PCA |
|  | **DETAILS ABOUT MACHINE LEARNING TECHNIQUES USED** |
| Common steps deployed for each technique | * Parameters provided are:   + Hyper-parameters to seek an optimal value   + number of folds to be used for cross-validation * Changing Learning rate * Checked model for overfitting and underfitting |
| Technique 1 | SVC  Precision of Only 50% |
| Technique 2 | logistic regression  Precision 87.53% |
| Technique 3 | KNN  Precision 79.5% |
| Technique 4 | Decision Tree  Precision 85.32% |
| Technique 5 | Random Forest  Precision 90.44% |
| Technique 6 | Adaboost  Precision 53.32% |
| Technique 7 | Gradient Boosting  Precision 90.86% |
| Technique 8 | Use the VotingClassifier method  Precision: 90.30% |
| Recommendations for future work, if any | With regards to Machine Learning model   * Optimization in keyword generation * Use other classification techniques * Exploring other hyper-parameter tuning   With regards to overall functioning and performance   * Improve execution time and performance * Comparison of results across R and Python * Apply Deep Learning techniques for better results * Visualization tool can be integrated for better user friendly interface |
|  |  |
| Final presentation prepared and uploaded to mUniversity | Date: 11 July 2020 |
| Code uploaded to mUniversity | Date: 11 July 2020 |
| Data uploaded to mUniversity | Date: 11 July 2020  (If data size is very large, coordinate with Aegis Class Schedule, Aegis IT) |
| Presentation Video uploaded to mUniversity | Date: 11 July 2020 |
|  |  |