Name: Himani Aryan
Email: himani1689@gmail.com
Country: United Kingdom
University: University of Liverpool

TASK: INTERSNHIP WEEK 8

SPECIALITY: DATA SCIENCE

1. **Problem Description:**

ABC Pharma wants to automate the identification process and understand the persistence of drug use based on medical prescriptions. In order to predict patient persistence in taking prescribed medications, they seek to develop a classification model. Better patient outcomes, higher sales, and lower healthcare costs can all result from improved drug persistency.

2. **Data Understanding:**

The National Institute of Diabetes and Digestive and Kidney Diseases is the source of the medical data in the Diabetes Dataset. The main goal of this dataset is to predict a patient's diagnosis of diabetes based on specific diagnostic parameters.

- Dataset Name: Diabetes Dataset
- Dataset Link: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

3. **Key points about the dataset:**

↔ **Data Source:** The dataset was sourced from the National Institute of Diabetes and Digestive and Kidney Diseases.

↔ **Target Variable**: The dataset includes a target variable called "Outcome." This variable is binary, with values 0 and 1, representing the absence or presence of diabetes, respectively.

↔ **Patients' Characteristics**: All patients in the dataset are females, at least 21 years old, and belong to the Pima Indian heritage. This information is essential to understand the demographic profile of the patients in the dataset.
This dataset contains medical claims data related to diabetes patients, which can be relevant for analyzing drug persistency. It includes various features such as patient DiabetesPedigreeFunction, Age, BMI, Insulin etc.

↔ **Independent Variables**: The dataset contains several independent variables, which are medical predictor variables used to diagnose diabetes. These variables could include measurements such as blood glucose levels, insulin levels, body mass index (BMI), age, and other relevant medical factors.

↔ **Data Format**: The dataset is provided in a (.csv) file format

- Machine learning algorithms will be applied to this dataset to create a classification model that can predict whether a patient has diabetes or not, given the values of the independent variables.
- This dataset will be used to explore patterns and relationships between different medical predictors and the occurrence of diabetes.

- While the dataset is focused on diabetes patients, it can serve as a good starting point for understanding drug persistency in the context of chronic diseases.
- Pre-processing of dataset will be done to suit the specific requirements of predicting drug persistency for ABC Pharma.

4. Problems in the Data:

Based on specific diagnostic parameters, the major goal of this dataset is to diagnostically predict whether a patient has diabetes or not.

It may also present some challenges and issues that should be taken care of.

Following are some potential problems that might be encountered when working with this dataset:

↔ Imbalanced Data: The dataset can have a class imbalance, which occurs when one class (for example, patients without diabetes) is much more common than the other class (for example, patients with diabetes). Incorrect model performance and trouble correctly forecasting the minority class can result from unbalanced data.

↔ Missing Values: Improper imputation may introduce biases in the data analysis due to missing values.

↔ Outliers: Outliers will influence the model's learning process and should be appropriately addressed during data pre-processing phase.

↔ Feature Selection: To prevent overfitting and enhance model interpretability, it is crucial to choose the most important characteristics for the predictive model.

↔ Data Overfitting: It is possible that the model may overfit the training data, performing well on the training set but poorly on unseen data. Regularization techniques and cross-validation should be employed to mitigate overfitting.

↔ Confounding Factors: Identifying and accounting for confounding factors can be challenging but crucial for accurate predictions in this case.

↔ Data Privacy and Ethics: Medical data should be handled with utmost care to ensure patient privacy and comply with data protection regulations. Data sharing and usage should follow ethical guidelines to prevent potential harm to patients.

↔ External Validity: The dataset may have limitations in its representativeness, as it specifically includes females of Pima Indian heritage. The model's performance on other demographic groups or populations might be different, affecting its external validity.

5. Approaches to Overcome Problems:

To overcome the problems ,following techniques can be implemented:

↔ Imbalanced Data:
- Some techniques to balance the dataset such as oversampling the minority class (patients with low persistency) or under sampling the majority class (patients with high persistency) can be applied.

↔ Missing Values:
- Thorough data cleaning and handle missing values appropriately using techniques like mean, median, or mode imputation based on the type of data can be done.
- Advanced imputation techniques like KNN or model-based imputation can be used.

↔ Outliers:
- Detection and handling of outliers using techniques like Z-score or Interquartile Range (IQR) method can be done.
- Later, evaluation of the impact of outliers on the model can be done and data transformation techniques can be applied accordingly.

↔ Feature Selection:
- Utilization of techniques L1 regularization (Lasso) can be done to select the most important features and reduce overfitting.

↔ Data Overfitting:
- Regularization techniques such as L1 or L2 regularization can be applied to prevent overfitting in the model.
- Cross-validation to can be opted to evaluate model performance on different subsets of the data and ensure generalization.

↔ Confounding Factors:
- Implementation of statistical methods like propensity score matching to control for confounding effects can be applied, if needed.

- Additionally, during the EDA phase, exploration of the relationship between variables and the target variable (persistency) should be done.
- Identification of any potential biases or confounding factors should be taken care of.

6. Implementation and Monitoring:

Implementation and monitoring are crucial phases in the project lifecycle, ensuring that the insights and recommendations derived from the classification model for drug persistency

prediction are effectively put into action and continuously assessed for their impact. Here's a detailed explanation of these phases:

Implementation:

↔ **Model Deployment**: The trained and validated classification model should be deployed into ABC Pharma's systems or applications to enable real-time predictions. Integration with the existing infrastructure ensures that the model can be utilized in day-to-day operations.

↔ **Automation:** Automate the identification process using the classification model to predict drug persistency based on doctor prescriptions. This automation streamlines the identification of patients with higher persistency levels and improves efficiency.

↔ **Customized Marketing Strategies**: Implement customized marketing strategies based on the insights from the model. Target specific patient groups with tailored campaigns to promote medication adherence and engagement.

↔ **Optimized Treatment Programs**: Use the model's predictions to optimize treatment plans based on factors influencing drug persistency. Tailor treatment programs to enhance patient adherence to prescribed medications and treatment regimens.

↔ **Stakeholder Training**: Train relevant stakeholders, including healthcare providers, marketing teams, and decision-makers, on how to interpret and utilize the model's predictions effectively.

↔ **Performance Metrics:** Define clear performance metrics for the implemented strategies, such as patient adherence rates, revenue trends, and healthcare expenses. These metrics will help in measuring the effectiveness of the implemented changes.

Monitoring:

↔ **Continuous Assessment**: Regularly monitoring the performance of the implemented strategies and the classification model's predictions.
Continuously assessing the impact of the strategies on patient outcomes, revenue, and healthcare expenses.

↔ **Model Performance:** Monitoring the model's performance metrics, including accuracy, precision, recall, and F1-score.
Keeping track of any changes in model performance over time.

↔ **Model Maintenance**: Updating the model periodically to adapt to any changes in patient demographics, treatment guidelines, or medical practices.
Retraining the model with new data could also be done to ensure its relevance and accuracy.

↔ **Data Quality Checks**: Regularly conducting data quality checks to ensure that the data used for prediction remains reliable and relevant.
Addressing any issues related to missing data or data discrepancies.

↔ **Review of Recommendations**: Revisiting the recommendations provided based on the model's predictions and verifying their effectiveness in improving drug persistency.

↔ **Stakeholder Engagement**: Maintaining open communication with stakeholders from ABC Pharma and involving them in the monitoring process.

- By implementing the model's predictions and closely monitoring its performance and impact, ABC Pharma can continuously optimize their marketing strategies, treatment programs, and patient outcomes.
- This iterative process allows for data-driven decision-making, leading to improved drug persistency and overall success for the organization.

7. Ethics and Data Privacy:

- ABC Pharma should ensure that the data is used solely for the intended purposes and that the results are reported in an aggregated and non-identifiable manner.

- Any potential biases in the data or models should be carefully addressed to ensure fair and equitable decision-making.

- Transparency and ethical considerations in the use of AI and machine learning algorithms are critical to maintaining trust with patients and healthcare professionals.

In conclusion, by following a systematic project lifecycle, conducting thorough data analysis, and implementing actionable recommendations, ABC Pharma can gain valuable insights into drug persistency and optimize their strategies to improve patient outcomes, revenue, and overall healthcare efficiency.