DSCI 542: Lab3 Question4

Block 3 Group 15

2025-01-26

Talk outline + Slide submission

rubric: {presentations + outline rubric = 10+3}

Talk Outline

Importance of Diabetic Study [Inder]

- **Widespread Condition**:
  - Discuss the increasing rates of both type 1 and type 2 diabetes
  - Chronic nature of diabetes can have lifelong impact on individuals and societies.
- **Identifying Predictors for Preventive Measures**:
  - Importance of early diagnosis and understanding risk factors.
  - Can reduce Healthcare Costs and Improving Quality of Life

Previous Research [Inder]

- **Neural Networks**:
  - Strengths: ability to capture complex relationships in data
  - Limitations: model interpretability & the need for large datasets.
- **Logistic Regression**:
  - Strength: Simplicity, efficiency, and good performance
  - Limitation: Outliers in data effect model
- **Random Forest**:
  - Strengths: Good at dealing with missing data, and less prone to overfitting
  - Limitations: Less interpretability, overly complex model

What makes our study different [Inder]

- Explain the importance of data validation techniques, ensuring robust and reliable results.
- Discuss strategies for handling class imbalance
- Use Logistic Regression to interpret coefficients

Reason of Choosing Logistic Regression [Inder]

- Study designed to predict if patient is diabetic or not which is Binary Classification
- Coefficients tell you the strength and direction of each feature's impact on the outcome
  - (e.g., which factors increase or decrease the likelihood of diabetes)
- Probabilities provide richer information, such as a probability score that quantifies the likelihood of an individual having diabetes

Dataset - Scope [Javier]

- Dataset origin (National Institute of Diabetes and Digestive and Kidney Diseases)
- Patient demographics (768 female patients, Pima Indian heritage)
- Goal: Predict diabetes
- Data source: Kaggle

Dataset - Feature Descriptions [Javier]

- Table of dataset features (Pregnancies, Glucose, Blood Pressure, etc.)

Dataset - Class Imbalance [Javier]

- Explanation of imbalance (500 non-diabetic, 268 diabetic)
- Visualization of class distribution (bar chart)
- Challenges & strategies to address imbalance

Validation - Pandera [Inder]

- Need to validate dataset to get rid of outliers
  - Features had values that were not medically possible i.e. (Gluvose being 0)
- Previous study using Logistic Regression failed to do this
  - should increase the accuracy for our model
- Justification of values chosen for each feature

Validation - Deepchecks [Inder]

- Looking at correlation table see that Age & Pregnancy have Pearon correlation score of 0.56
  - Close to Multicollinearity threshold of 0.7
  - Multicollinearity can lead to unstable model coefficients, overfitting, and reduced interpretability
  - Explain how to use Deepchecks to confirm that feature do not pass correlation threshold

Analysis Methodology [Jenny]

- **Data Split**:
  - 70% for training, 30% for testing to evaluate model performance.
  - Ensures robust training while testing model generalization.
- **Features**:
  - Structured numeric data with no missing values—ensuring the dataset is clean and ready for modeling.
- **Preprocessing**:
  - Standardization applied to all features for consistent scale.
  - Ensures stable model performance and easy interpretation of coefficients.

Analysis Methodology (cont'd) [Jenny]

- **Evaluation Metric**:
  - Accuracy used to measure model performance.
  - Provides a straightforward metric to assess how well the model predicts diabetes.
- **Baseline Model**:

  - DummyClassifier serves as the baseline, providing a simple comparison.

- **Logistic Regression**:
  - Chosen for its ability to handle classification tasks effectively.
  - Its coefficients help identify the importance of each feature, offering insight into how each variable influences the model's predictions.
  - Provides probability-based predictions for interpretability.

- **Hyperparameter Tuning**:
  - RandomizedSearchCV used to optimize `C` (model complexity).
  - The range of `C` spans from 10^{-5} to 10^{5} to find the best balance between regularization and data fitting.

Analysis - Reproducible Data Pipeline [Javier]

- How a reproducible data pipeline was built
- Custom functions for data preprocessing, feature selection, training, and evaluation
- Outputs stored as tables, plots, models (pickle)
- Dynamic parameter tuning and modular workflow

Analysis - Modular Function Design [Javier]

- Automation & Interdependency of functions
- Code snippet: Saving models with pickle
- Ensuring traceability and reproducibility

Results - Feature Importance [Jenny]

- **Feature Importance via Coefficients**:
  - Coefficients represent the influence of each feature on the model's predictions.
  - **Glucose** is the strongest positive influence (0.724), followed by **BMI** (0.389).
  - **Pregnancies**, **Age**, and **DiabetesPedigreeFunction** also contribute but with smaller effects.
  - **SkinThickness** has a negative influence (-0.007), indicating minimal impact.

Results - Model Evaluation [Jenny]

- **Baseline Model**:
  - Accuracy of **DummyClassifier**: 0.672, providing a reference for comparison.
- **Logistic Regression Performance**:
  - Training accuracy: **0.743** (cross-validation mean).
  - Test accuracy: **0.750** - indicating reasonable model generalization.
  - The accuracy is solid but leaves room for improvements, especially considering clinical applications.

Results - Confusion Matrix [Jenny]

- **Misclassifications**:
  - Out of 217 total test cases, 54 misclassifications.
  - **41 false negatives**: Diabetic cases predicted as non-diabetic—critical in clinical settings.
  - **13 false positives**: Non-diabetic cases predicted as diabetic—less critical but still a concern.
  - In a clinical context, reducing **false negatives** is crucial for patient safety and intervention.

Results - PR and ROC Curve [Jenny]

- **Precision-Recall Curve**:
  - Evaluates the trade-off between true positives and false positives at various thresholds.
  - **No optimal threshold** observed that balances both high precision and recall.
- **ROC Curve**:
  - Shows the model's true positive rate vs. false positive rate.
  - Similarly, **no optimal threshold** observed that balances both high true positive and low false positive rates.
  - Further model adjustments may be needed to improve performance across all thresholds.

Results - Clinical Utility [Jenny]

- **Visualization of Predicted Probabilities**:
  - Helps clinicians understand the model's confidence in its predictions.
  - If probability is not high enough, additional tests may be considered.
  - This visualization aids in clinical decision-making, highlighting both correct predictions and false negatives.

Results - Clinical Utility [Jenny]

- **Visualization of Predicted Probabilities**:
  - Helps clinicians understand the model's confidence in its predictions.
  - If probability is not high enough, additional tests may be considered.
  - This visualization aids in clinical decision-making, highlighting both correct predictions and false negatives.

Discussion - Model Performance [Jessica]

- Clinical Relevance:
  - Model demonstrates potential as an initial screening tool for diabetes detection, especially given its improvement over the baseline.
  - Offers data-driven support for identifying at-risk individuals.
- Enhancement Approaches:
  - Examine 54 misclassified observations and compare with correctly classified examples to identify feature contributions.
  - Example: Combine related features or create new derived metrics (e.g., BMI adjusted for age).

Discussion - Enhancement Opportunities [Jessica]

- Explore feature engineering to improve model predictions.
- Experiment with alternative classifiers:
  - Random Forest: Accounts for feature interactions automatically.
  - k-Nearest Neighbours (k-NN): Offers interpretable and decent predictions.
  - Support Vector Classifier (SVC): Enables non-linear prediction using rbf kernel.

Limitations & Future Directions [Jessica]

- Dataset Limitations: Focuses solely on Pima Indian women aged 21 and older, limiting generalizability.
- Future Data Exploration:
  - Collaborate with data collectors for additional, useable information.
  - Combine with external datasets:
    * Broaden demographic coverage (age, gender, ethnicity)
    * Enable comprehensive insights and greater applicability.
    * Broaden demographic representation (diverse age groups, genders, and ethnicities).

Conclusion - Key Findings [Jessica]

- Logistic regression effectively predicts diabetes among Pima Indian women using features like glucose, BMI, and pregnancies.
- Achieved 0.750 accuracy on the test set, outperforming baseline Dummy Classifier's 0.672.
- Key predictors: Glucose (most influential) BMI, pregnancies.
- Challenges: 54 misclassifications (41 false negatives) present risks of undiagnosed cases, underscoring areas for refinement.

Conclusion - Clinical Implications [Jessica]

- Clinical Potential:
  - Logistic regression shows promise as an initial screening tool for early diabetes detection.
  - Provides a data-driven approach to improve outcomes and reduce complications.

Conclusion - Recommendations [Jessica]

- Recommendations for Improvement:
  - Model Improvement:
    * Refine predictors through feature engineering and additional data exploration.
    * Experiment with alternative machine learning models for better performance.
  - Expand Dataset:
    * Incorporate additional data sources, such as lifestyle or genetic information, to improve accuracy and generalizability.
  - Decision-Making Aid:

* Include probability estimates to support clinicians in identifying high-risk cases requiring further diagnostics.