INTEGRATED ROBUST SOLUTIONS, ISLAMABAD

Week 2 Task: AI Internship

Data Handling and Practical ML Tasks

Prepared by: Asad Ali
Designation: Lead AI Developer
Department: Artificial Intelligence

Version: 1.0

Date: July 11, 2025

Prepared for IR Solutions AI Interns

Contents

1 Introduction

This document outlines the Week 2 tasks for AI interns at IR Solutions. The tasks focus on building expertise in handling machine learning (ML) data through theoretical learning and practical implementation. Interns will study selected modules from Google's Machine Learning Crash Course and complete two programming tasks in Python using Google Colab: customer churn prediction and heart disease prediction, both using logistic regression. Weekly deliverables include a PDF report and Colab .ipynb files, submitted by **Friday, July 18, 2025, 5:30 PM PKT**.

2 Objectives

The Week 2 tasks aim to:

- Develop skills in handling numerical and categorical data for ML models.
- Understand dataset characteristics, generalization, and overfitting prevention.
- Apply data preprocessing and logistic regression to real-world ML problems.
- Practice professional documentation through structured PDF reports and well-commented Colab notebooks.
- Adhere to best programming practices (e.g., PEP 8, modular code, clear documentation).

3 Theoretical Learning

Interns are required to complete the following modules from Google's Machine Learning Crash Course (`https://developers.google.com/machine-learning/crash-course`):

3.1 Working with Numerical Data

- **Course Material**: Working with Numerical Data
- **Learning Objectives**:
  - Learn techniques to analyze and transform numerical data (e.g., normalization, scaling).
  - Understand the impact of numerical data preprocessing on model performance.

3.2 Working with Categorical Data

- **Course Material**: Working with Categorical Data
- **Learning Objectives**:
  - Distinguish categorical data from numerical data.
  - Apply encoding techniques (e.g., one-hot encoding, feature hashing, mean encoding) and feature crosses.

3.3 Datasets, Generalization, and Overfitting

- **Course Material**: Datasets, Generalization, and Overfitting
- **Learning Objectives**:
  - Understand dataset characteristics and their role in ML model quality.
  - Learn strategies to ensure generalization and prevent overfitting.

4 Practical Tasks

Interns will complete two programming tasks in Python using Google Colab. Each task includes a dataset and specific questions to guide implementation and analysis. Submit the Colab notebooks as `Week2_InternName_Task01.ipynb` and `Week2_InternName_Task02.ipynb`.

4.1 Task 01: Customer Churn Prediction using Logistic Regression

- **Description**: Build a logistic regression model to predict customer churn for a bank's savings account based on features like age, gender, dependents, occupation, and transaction data. The dataset (`churn_prediction.csv`) is available in the shared Google Drive folder.
- **Steps**:
  - Load and explore the dataset.
  - Handle missing values and preprocess numerical and categorical data.
  - Train a logistic regression model using scikit-learn.
  - Evaluate the model using AUC-ROC, recall, and confusion matrix.
  - Generate visualizations (e.g., ROC curve, confusion matrix).
- **Questions**:
  1. Load the dataset using `pd.read_csv`. How many rows and columns are in the dataset? List the column names.
  2. Identify missing values. How were missing values handled for `gender`, `dependents`, `occupation`, and `city`?
  3. Apply one-hot encoding to the `occupation` column. Why is one-hot encoding preferred over label encoding for this feature?
  4. Use `StandardScaler` and log transformation for numerical columns. Explain why log transformation was applied to balance features.
  5. Split the dataset into training (2/3) and validation (1/3) sets with `stratify = y_all`. Why is stratification important?
  6. Train a logistic regression model on the baseline features. List the selected features and explain their relevance based on EDA insights.
  7. Plot the ROC curve and report the AUC score. What does the AUC indicate about model performance?
  8. Generate and interpret the confusion matrix. What are the implications of the recall score?
  9. Discuss how overfitting was addressed in the model. Suggest one method to improve the recall score.

10. Save the model using `pickle`. Why is this step important for reproducibility?

- **Deliverables**:
  - Colab notebook (`Week2_InternName_Task01.ipynb`) with answers to the questions in Markdown cells, complete code, and visualizations (e.g., ROC curve, confusion matrix).
  - Include comments following PEP 8 standards and ensure the notebook is executable.

4.2 Task 02: Predicting Heart Disease using Machine Learning

- **Description**: Build a logistic regression model to predict heart disease based on medical features (e.g., age, sex, cholesterol, chest pain type). Compare its performance with KNN and Random Forest models. The dataset (`heart-disease.csv`) is available in the shared Google Drive folder.

- **Steps**:
  - Load and explore the dataset.
  - Preprocess numerical and categorical data.
  - Train and compare logistic regression, KNN, and Random Forest models using scikit-learn.
  - Perform hyperparameter tuning using RandomizedSearchCV or GridSearchCV.
  - Evaluate models using accuracy, precision, recall, F1-score, and AUC; generate visualizations.

- **Questions**:
  1. Load the dataset using `pd.read_csv`. How many rows and columns are in the dataset? List the column names.
  2. Check for missing values. Why is it significant that this dataset has no missing values?
  3. Analyze the correlation matrix using `df.corr()`. Identify the top three features correlated with the target and explain their importance.
  4. Split the dataset into training (80%) and validation (20%) sets. What are the sizes of these sets?
  5. Train a logistic regression model with
  6. Compare the accuracy of logistic regression, KNN, and Random Forest models. Which model performed best, and why?
  7. Perform hyperparameter tuning for KNN. Plot train and test scores. What is the optimal number of neighbors?
  8. Generate the confusion matrix and classification report for the logistic regression model. Interpret the precision, recall, and F1-score.
  9. Perform 5-fold cross-validation for the logistic regression model. Report the mean accuracy, precision, recall, and F1-score.
  10. Visualize feature importance for the logistic regression model. Which feature has the highest impact, and why might this be significant?

- **Deliverables**:

  - Colab notebook (`Week2_InternName_Task02.ipynb`) with answers to the questions in Markdown cells, complete code, and visualizations (e.g., ROC curve, confusion matrix, feature importance).
  - Include comments following PEP 8 standards and ensure the notebook is executable.

5 Submission Guidelines

5.1 PDF Report

- **Format**:
  - **Title Page**: Include IR Solutions logo, intern's name, designation (AI Intern), department (Artificial Intelligence), and date.
  - **Sections**:
    - **Theoretical Understanding**: Summarize key concepts from the assigned course modules.
    - **Project Work**: Describe the two tasks, including methodology, code implementation, and results.
    - **Challenges and Approaches**: List challenges faced in the tasks and solutions applied.
    - **Results and Analysis**: Present model performance (e.g., AUC, recall, accuracy) with plots (e.g., ROC curves, confusion matrices).
    - **Spare Time Activities**: Note any additional learning or exploration during spare time.
    - **Conclusion**: Summarize key takeaways and plans for improvement.
    - **References**: Cite all resources used.
  - Use professional formatting with headings, subheadings, bullet points, and tables/plots for results.
- **File Naming**: `Week2_InternName_Report.pdf`

5.2 Colab Notebooks

- Submit two notebooks: `Week2_InternName_Task01.ipynb` and `Week2_InternName_Task02.i`
- Follow PEP 8 standards, include clear comments, and use print statements for debugging.
- Use Markdown cells for task overview, methodology, question answers, and hyperparameter analysis.
- Include visualizations (e.g., ROC curves, confusion matrices, feature importance plots).
- Ensure notebooks are fully executable with no errors.

5.3 Submission Process

- Upload the PDF report and both .ipynb files to the designated Google Drive folder by **Friday, July 18, 2025, 5:30 PM PKT**.
- Ensure files are clearly named and organized in the Week 2 folder.

6 Evaluation Criteria

Interns will be evaluated based on:

- **Theoretical Understanding (30%)**: Depth and clarity of concepts learned from the course modules.
- **Project Output & Code Implementation (30%)**: Correctness, efficiency, and documentation of code; quality of analysis and visualizations.
- **Report Quality (20%)**: Clarity, completeness, and adherence to the defined format.
- **Best Practices (10%)**: Use of PEP 8, modular code, and clear documentation.
- **Participation (10%)**: Engagement in the Week 2 AI standup meeting and responsiveness to feedback.

7 Additional Notes

- **Dataset Access**: Both datasets (`churn_prediction.csv` and `heart-disease.csv`) are available in the shared Google Drive folder.
- **Support**: Reach out to AI team members or Lead AI via Slack or during the standup meeting (Friday, 3:00 PM PKT) for guidance.
- **Expectations**: Maintain professionalism, meet deadlines, and actively engage in learning and collaboration.

8 References

- Google Machine Learning Courses: `https://developers.google.com/machine-learning`
- GeeksforGeeks:
  - `https://www.geeksforgeeks.org/machine-learning`
  - `https://www.geeksforgeeks.org/machine-learning-projects`
  - `https://www.geeksforgeeks.org/deep-learning-tutorial`
  - `https://www.geeksforgeeks.org/computer-vision`
  - `https://www.geeksforgeeks.org/natural-language-processing-nlp-tutor`
- YouTube Tutorials:
  - `https://www.youtube.com/watch?v=_3y7C7_do-k`
  - `https://www.youtube.com/watch?v=7Bg3iBq-SIY`
  - `https://www.youtube.com/watch?v=0YdpwSYMY6I`
  - `https://www.youtube.com/watch?v=yLlgHKaMyIw`
  - `https://www.youtube.com/watch?v=NOJOYcmyDhM`