**SUMMER TRAINING**

**PROJECT REPORT**

(Term June-July 2025)

**Online Shopper Intention** **Prediction Model**

Submitted by

|  |  |
| --- | --- |
| Name | Registration Number |
| Rudra nayak | 12306470 |
| Nageshwar Tiwari | 12306394 |

**Course Code: PETV79**

Under the Guidance of

Mahipal Singh Papola (UID: 32137)

**Lovely Professional University ,Punjab**

**BONAFIDE CERTIFICATE**

This is to certify that the project report titled "Online Shopper Intention using Machine Learning" is a record of original work carried out by Rudra and Nageshwar during their summer internship as a part of their B.Tech (Computer Science and Engineering) curriculum at Lovely Professional University. The work has been completed under my guidance and supervision and is a part of their partial fulfillment of the degree.

SIGNATURE

Supervisor  
 Mahipal Sir  
 Department of CSE, LPU

RUDRA

NAGESHWAR

SIGNATURE(Students)

SIGNATURE

(HEAD OF DEPARTMENT)

**Acknowledgement**

We would like to express our heartfelt gratitude to our project guide Mahipal Sir for his invaluable support, continuous guidance, and encouragement throughout the duration of this project. We would also like to thank the faculty and staff of the School of Computer Science and Engineering, LPU, for providing the necessary academic environment and technical resources.

**TABLE OF CAPITALIZED**

1. **Introduction**1.1 Company Profile  
   1.2 Overview of Training Domain  
   1.3 Objective of the Project
2. **Training Overview**2.1 Tools & Technologies Used  
   2.2 Areas Covered During Training  
   2.3 Daily/Weekly Work Summary
3. **Project Details**3.1 Title of the Project  
   3.2 Problem Definition  
   3.3 Scope and Objectives  
   3.4 System Requirements
4. **Implementation**4.1 Tools Used  
   4.2 Methodology  
   4.3 Modules / Screenshots  
   4.4 Code Snippets
5. **Results and Discussion**5.1 Output / Report  
   5.2 Challenges Faced  
   5.3 Learnings
6. **Conclusion**6.1 Summary
7. **References**

**CHAPTER 1: INTRODUCTION**

1.1 COMPANY PROFILE

This project was undertaken as part of the academic curriculum at the School of Computer Science and Engineering, Lovely Professional University (LPU), Phagwara, Punjab. LPU is a premier educational institution in India, recognized for its focus on industry-oriented education and research. The School of Computer Science and Engineering is equipped with modern facilities and fosters innovation in fields such as machine learning, data science, artificial intelligence, and software engineering. The department emphasizes practical training through projects, enabling students to apply theoretical knowledge to real-world problems.

1.2 OVERVIEW OF TRAINING DOMAIN

The training domain for this project is Machine Learning (ML) and Data Science, with a focus on regression modeling, data preprocessing, exploratory data analysis (EDA), and web application deployment. The project involved developing a predictive model to estimate the shoppers intention online, a regression task requiring skills in data cleaning, feature engineering, model training, and user interface development. The training provided hands-on experience with Python-based ML tools and deployment frameworks, preparing us for real-world applications in the automotive industry.

1.3 OBJECTIVE OF THE PROJECT

The objectives of the project are:

* Preprocess session data from e-commerce logs.
* Train a classification model to predict purchase intention (Revenue).
* Build a Flask-based web application for real-time predictions.
* Evaluate model accuracy and propose improvements.

**CHAPTER 2: TRAINING OVERVIEW**

2.1 TOOLS & TECHNOLOGIES USED

The project utilized the following tools and technologies:

* PYTHON: Core programming language for data processing, modeling, and deployment.
* SCIKIT-LEARN: For implementing the Random Forest regression model and computing evaluation metrics.
* PANDAS: For data manipulation, cleaning, and preprocessing.
* NUMPY: For numerical computations and array operations.
* MATPLOTLIB AND SEABORN: For creating visualizations such as scatter plots, box plots, and regression plots.
* FLASK: For developing and deploying the interactive web application.
* JUPYTER NOTEBOOK: For exploratory data analysis, prototyping, and code development.

2.2 AREAS COVERED DURING TRAINING

The training covered the following key areas:

* DATA PREPROCESSING: Handling missing values, removing duplicates, and encoding categorical variables.
* EXPLORATORY DATA ANALYSIS: Performing univariate and bivariate analyses to identify patterns and correlations.
* FEATURE ENGINEERING: Extracting relevant features and cleaning numerical data.
* MODEL DEVELOPMENT: Training a Random Forest regression model and evaluating its performance.
* DEPLOYMENT: Building a user-friendly web application using Flask for real-time predictions.
* VISUALIZATION: Creating plots to visualize feature distributions and relationships.

2.3 DAILY WORK SUMMARY

The project was executed over five days, with the following milestones:

* Day 1: Dataset understanding and null value checks.
* Day 2: Data encoding, feature engineering.
* Day 3: EDA and visualizations.
* Day 4: Model training and validation.
* Day 5: Flask web app creation and testing.

**CHAPTER 3: PROJECT DETAILS**

3.1 TITLE OF THE PROJECT

Online Shopper Purchase Intention Prediction

3.2 PROBLEM DEFINITION

E-commerce platforms collect session data to understand user behavior. This project predicts whether a user will make a purchase based on these behavioral features, helping in targeted marketing and optimization.

3.3 SCOPE AND OBJECTIVES

SCOPE:

* Develop a machine learning model to predict online shoppers intention based on a dataset of 12,331.
* Ensure data quality through preprocessing and feature engineering.
* Deploy the model as a user-friendly web application for real-time predictions.
* Achieve high predictive accuracy (target R² > 0.85).

OBJECTIVES:

* Clean and preprocess the dataset to remove inconsistencies and ensure model compatibility.
* Engineer features to capture relevant information .
* Train a Random Forest model to predict with high accuracy.
* Deploy the model via Flask for accessibility to non-technical users.
* Evaluate model performance and identify areas for improvement.

3.4 SYSTEM REQUIREMENTS

* HARDWARE: Standard laptop/desktop with at least 8 GB RAM and a 2 GHz processor.
* SOFTWARE:
  + Python 3.8 or higher
  + Libraries: scikit-learn, pandas, NumPy, Matplotlib, Seaborn, Flask
  + Jupyter Notebook for development and testing
* DATASET: "online\_shoppers\_intention.csv"
* OPERATING SYSTEM: Windows, macOS, or Linux.

**CHAPTER 4: IMPLEMENTATION**

4.1 TOOLS USED

The implementation utilized:

* PYTHON LIBRARIES: pandas for data manipulation, NumPy for numerical operations, scikit-learn for model training and evaluation.
* VISUALIZATION TOOLS: Matplotlib and Seaborn for generating scatter plots, box plots, and regression plots.
* FLASK: For building and deploying the interactive web application.
* JUPYTER NOTEBOOK: For prototyping and exploratory analysis.

4.2 METHODOLOGY

The project followed a structured machine learning pipeline:

**Step 1: Data Collection**  
The dataset was collected from the UCI ML repository and consists of 12,331 sessions from an online shopping website. The data includes attributes such as page types, bounce rates, exit rates, time spent, and whether or not the session led to a transaction (Revenue).

**Step 2: Data Preprocessing**

* Missing values were checked and no nulls were found.
* The categorical columns Month, VisitorType, and Weekend were label-encoded.
* The target column Revenue was also converted to binary form.
* All numerical features were scaled using StandardScaler for uniformity.

**Step 3: Exploratory Data Analysis (EDA)**

* Count plots for Revenue distribution showed class imbalance.
* Correlation matrix showed strong influence from PageValues, BounceRates, and ExitRates.
* Box plots were created to understand spread across feature values.

**Step 4: Model Development**

* Data was split into training and testing using an 80:20 ratio.
* A RandomForestClassifier was trained due to its ability to handle mixed data types and prevent overfitting.
* Accuracy, precision, recall, and F1-score were calculated.

**Step 5: Web App Deployment Using Flask**

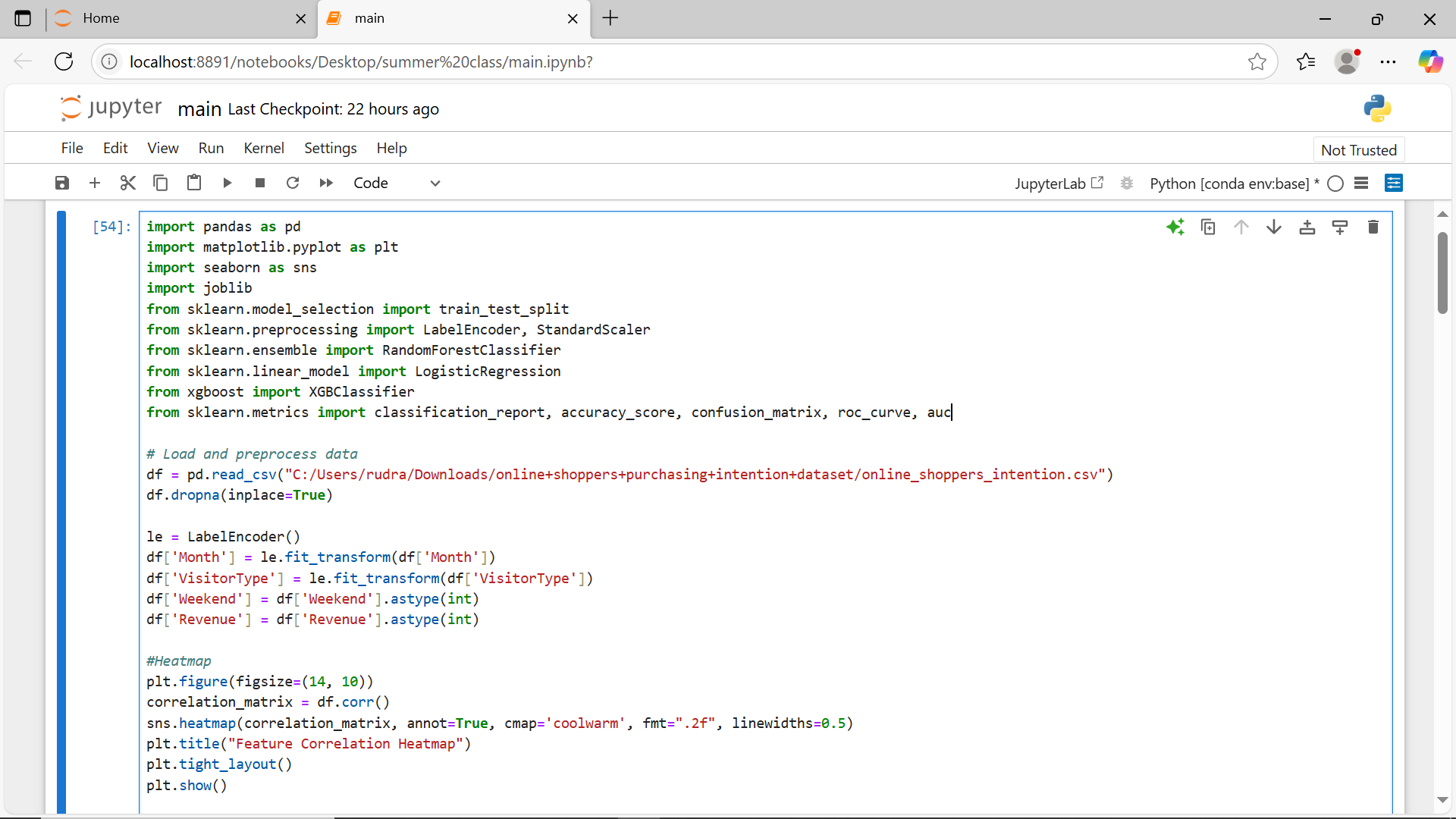
* A form-based web app was developed using Flask.
* Users input data like Month, VisitorType, Bounce Rate, Exit Rate, etc.
* The backend uses the saved Random Forest model to return prediction (purchase or not).

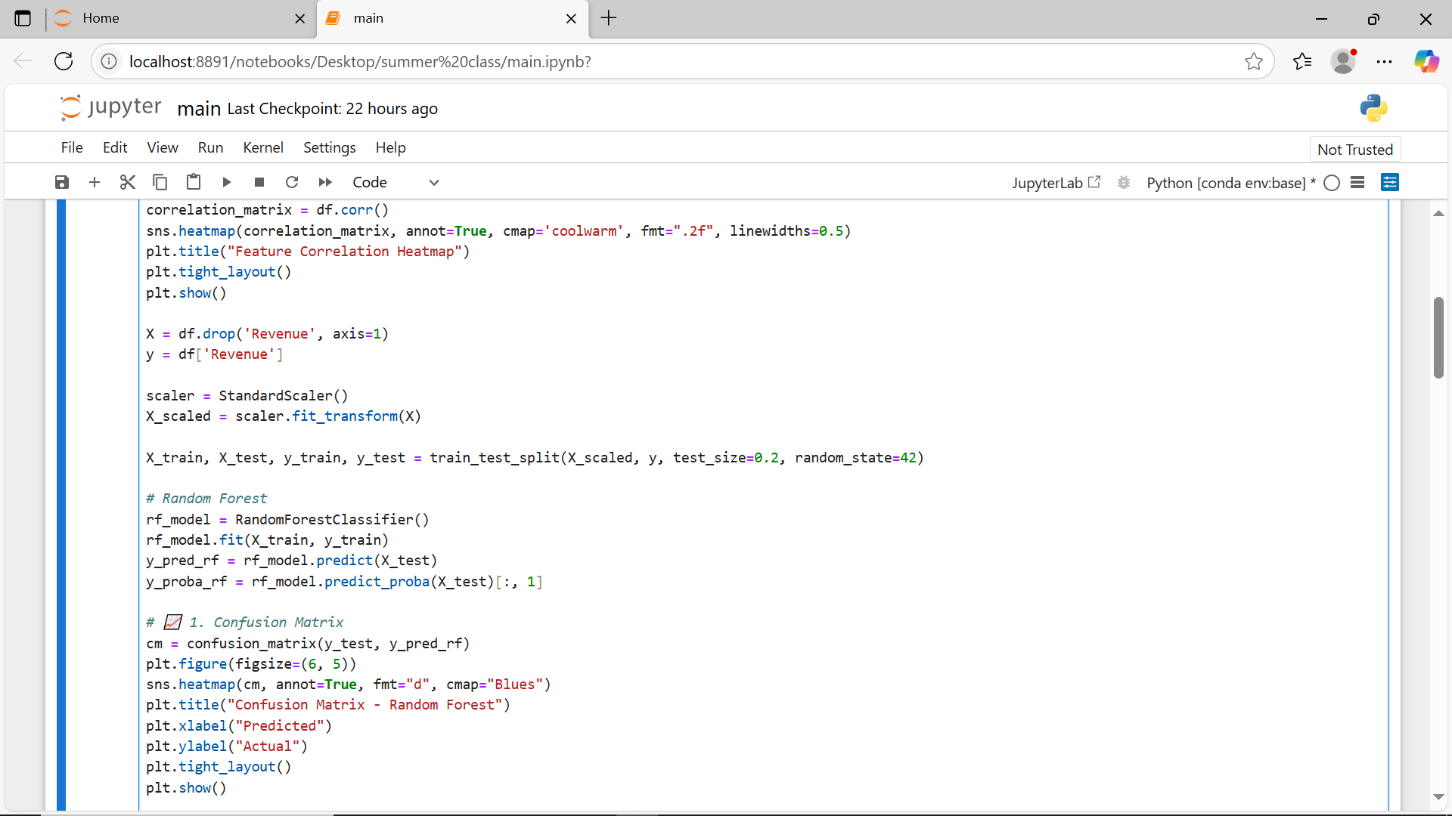
4.3 MODULES / SCREENSHOTS

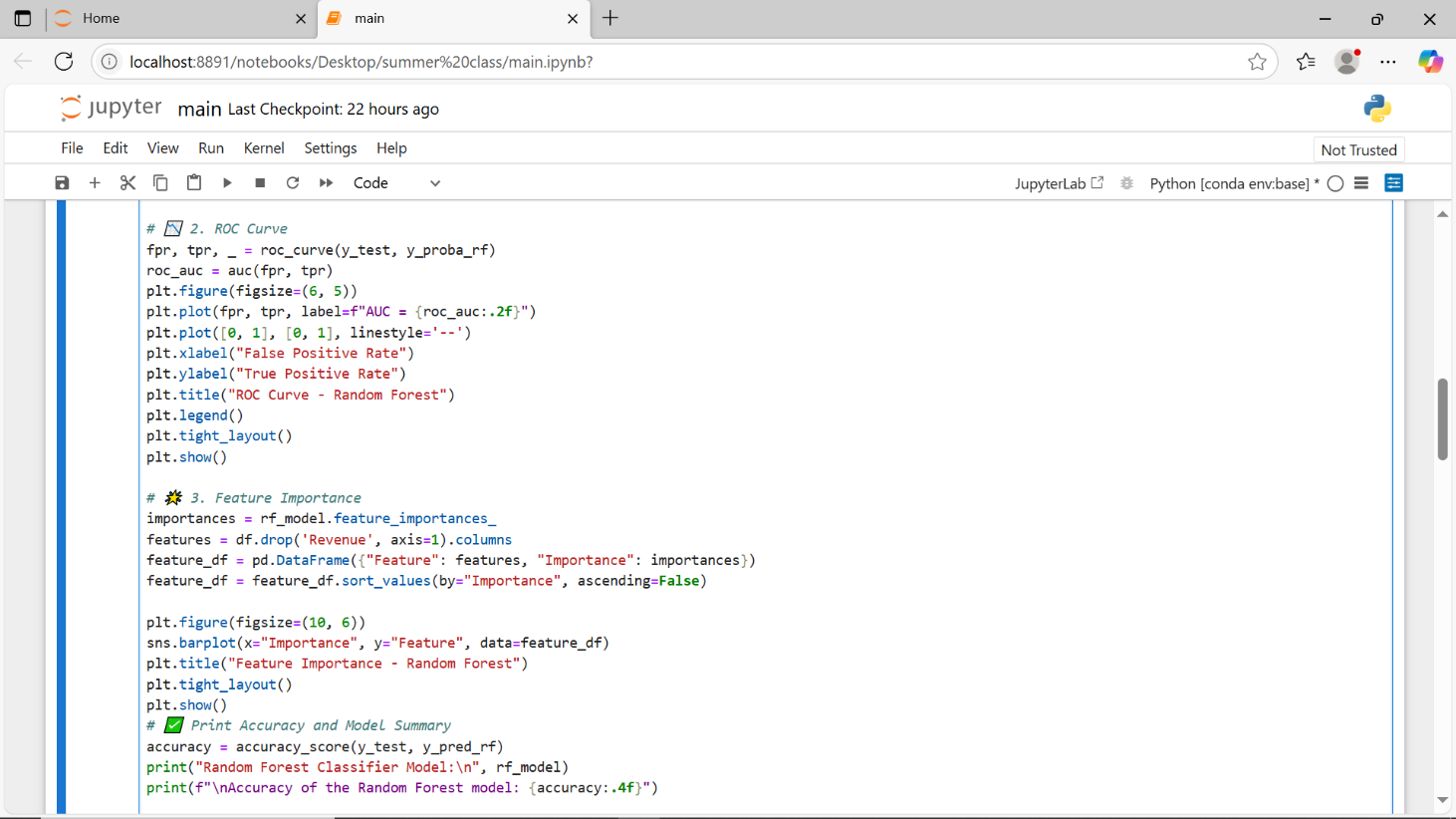
The project consists of the following modules:

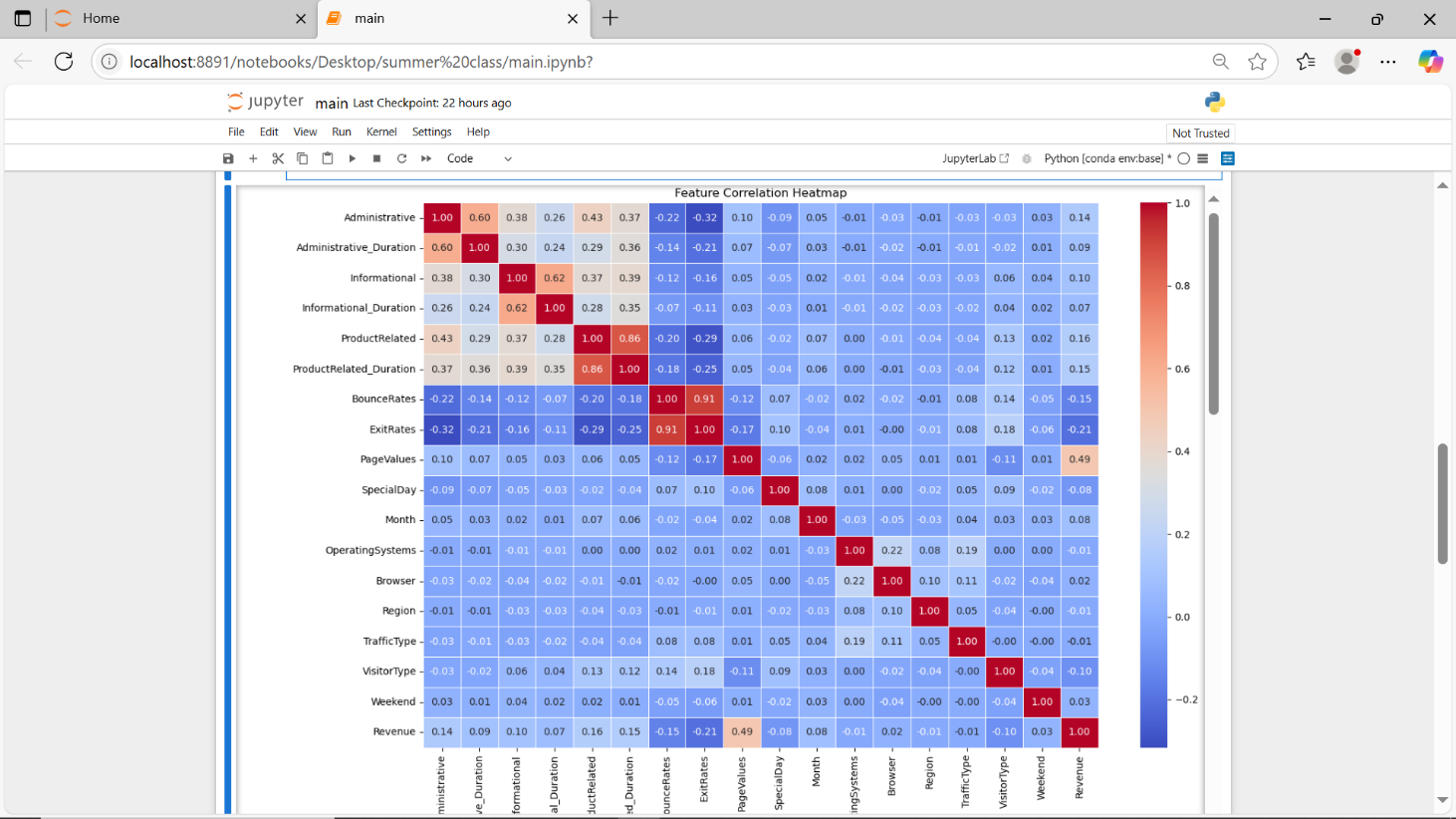
* **Preprocessing Module:** Reads data, encodes categorical values, scales features.
* **Model Module:** Trains and saves the Random Forest classifier.
* **Web Interface Module:** Flask app that allows user input and returns predictions.

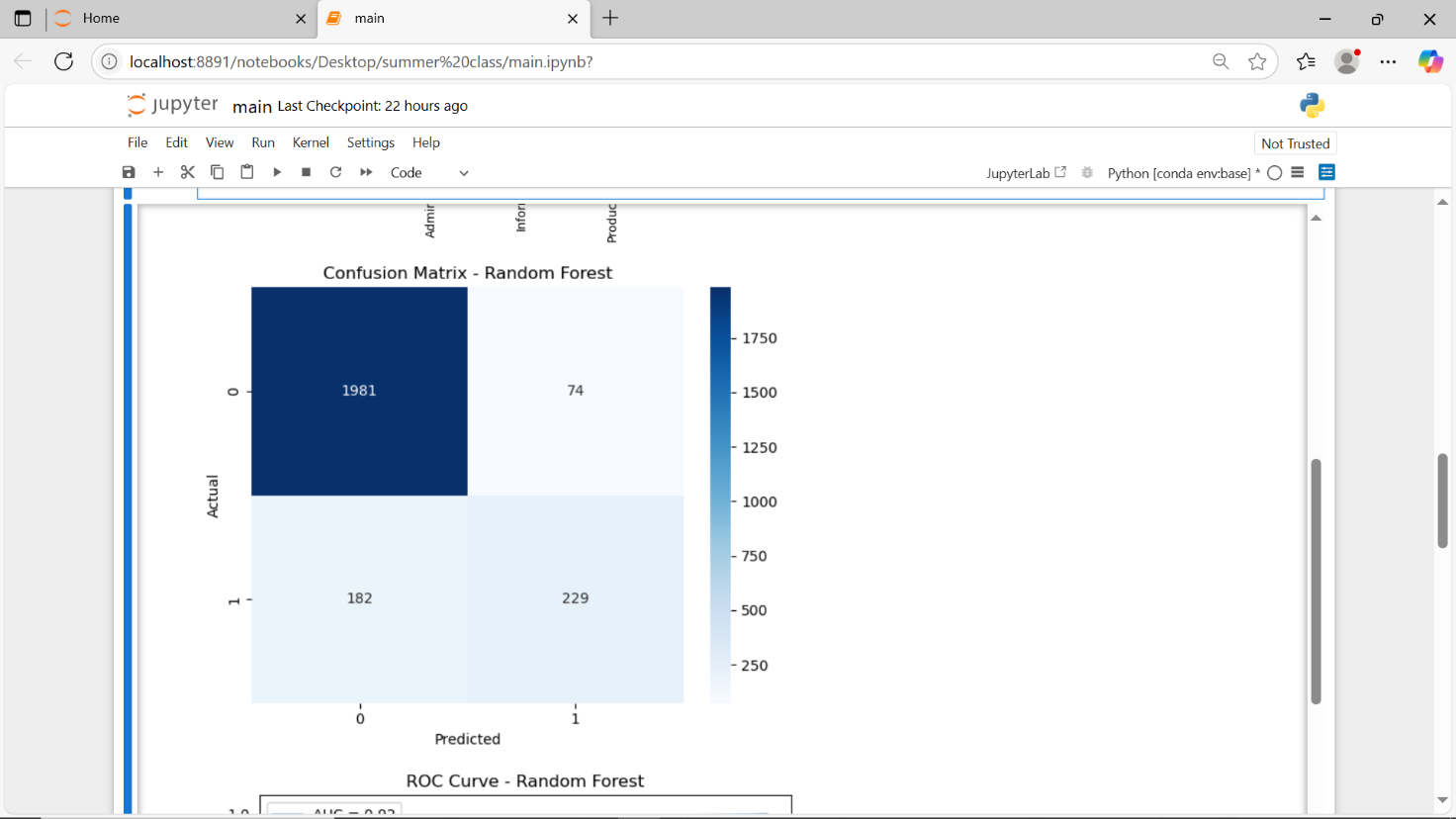
4.4 CODE SNIPPETS

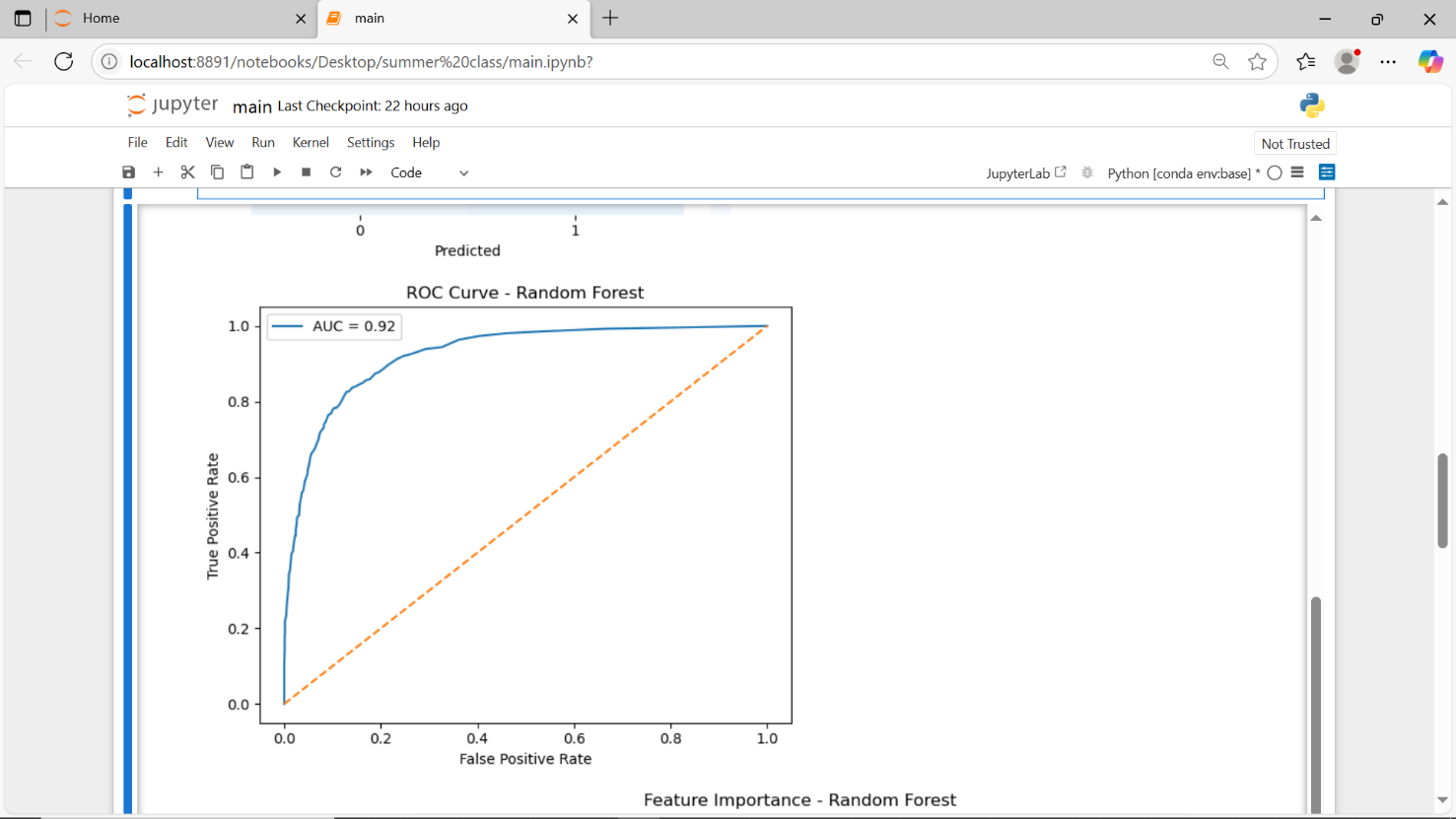


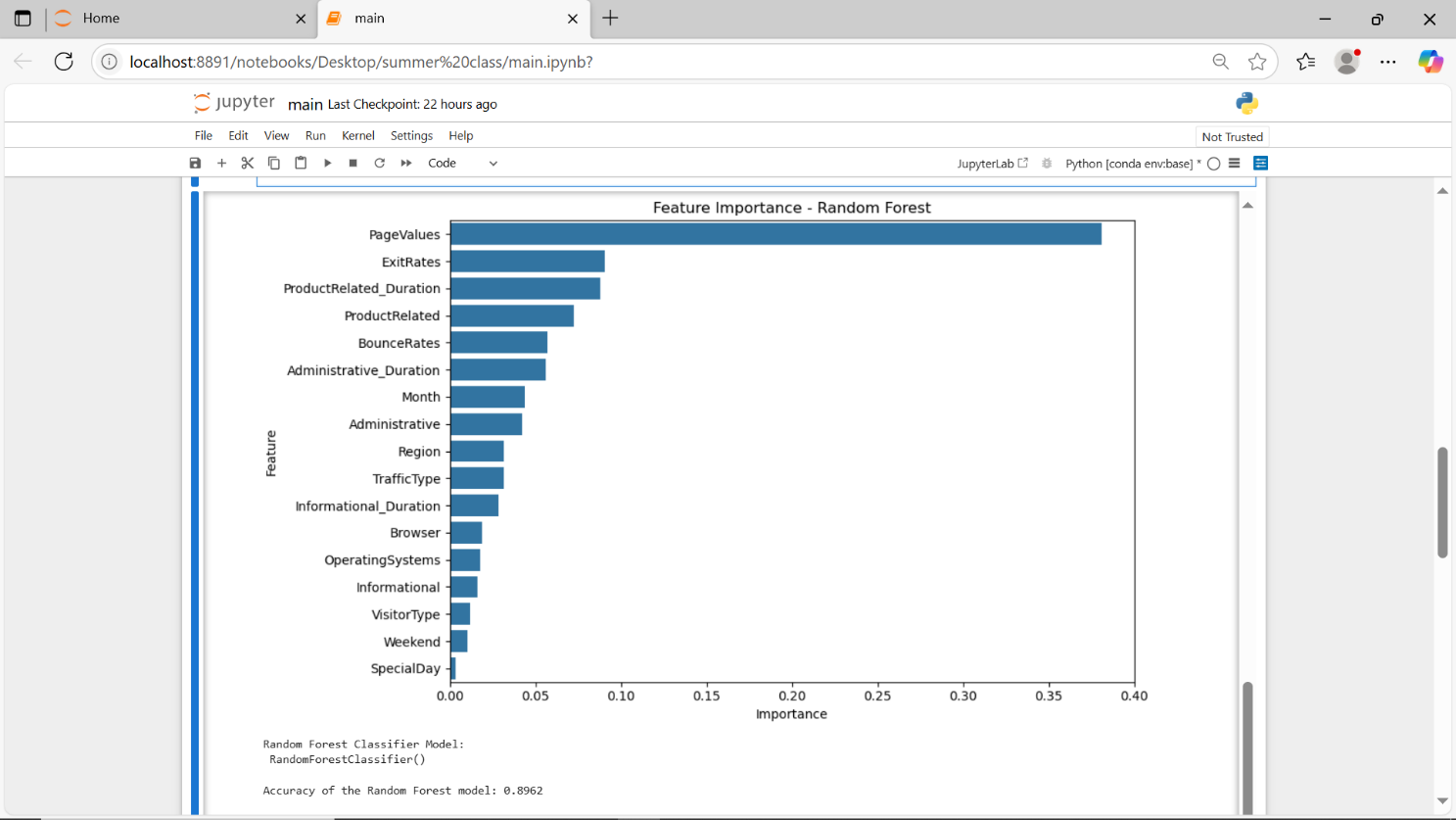


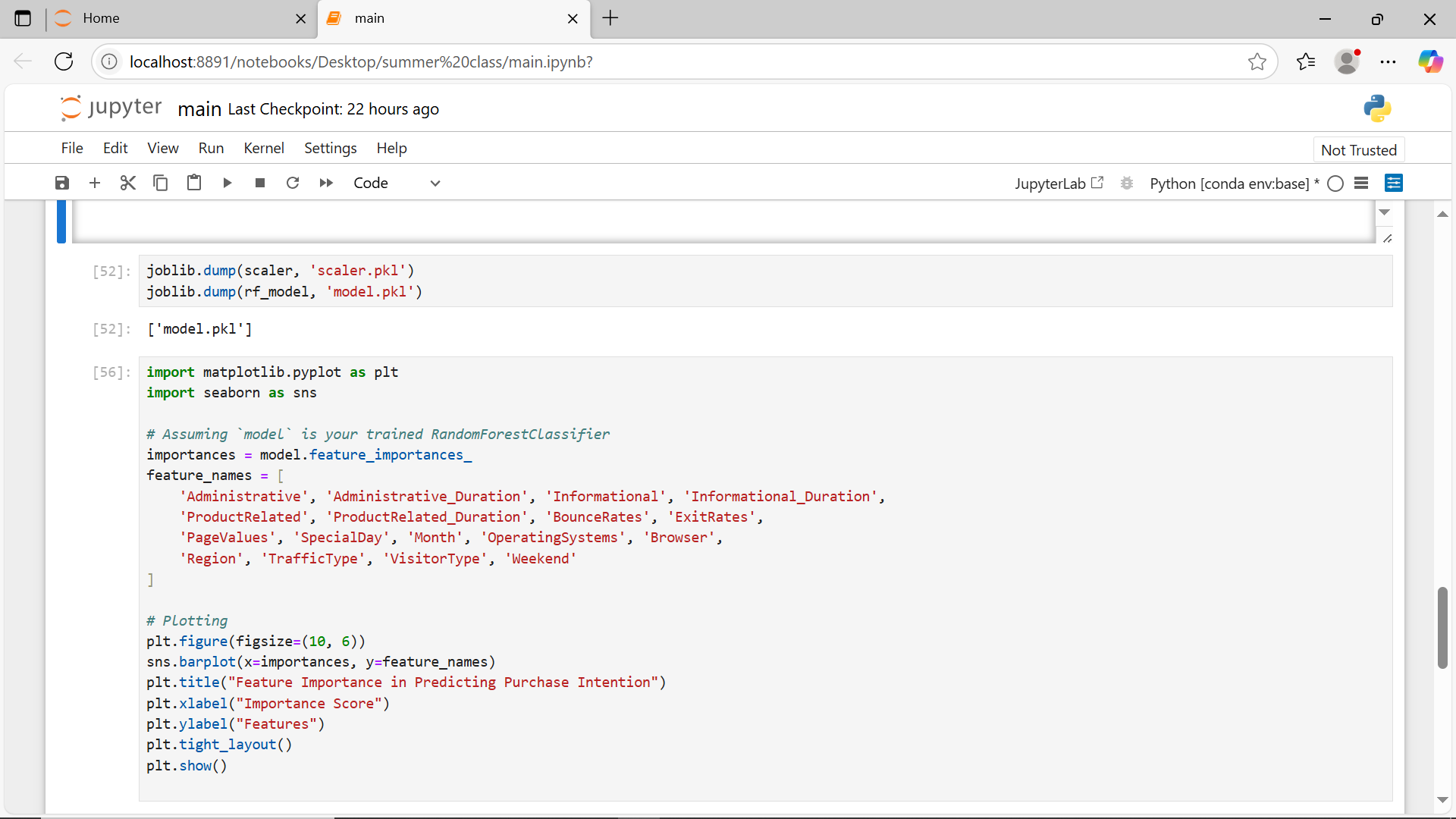


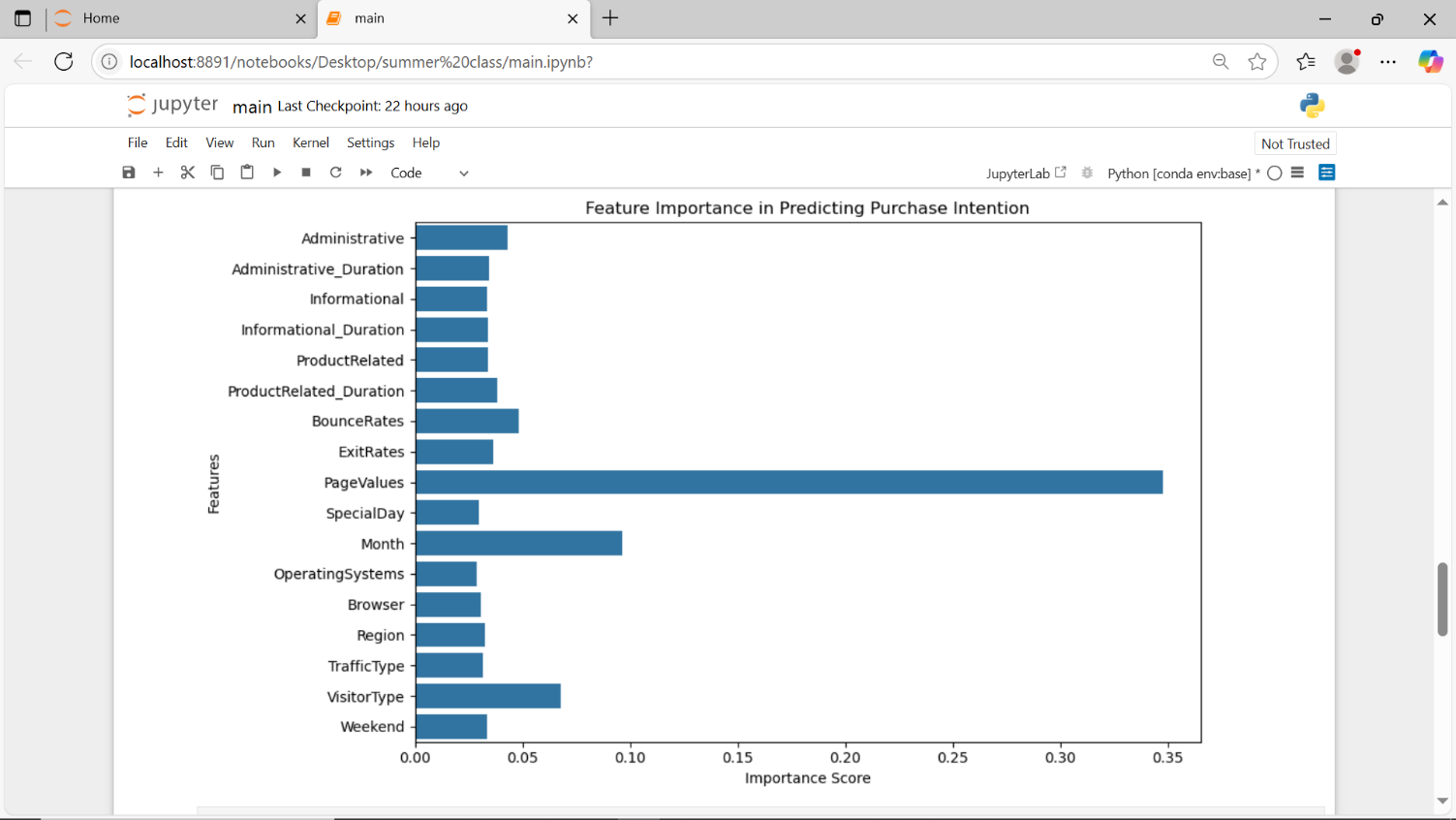


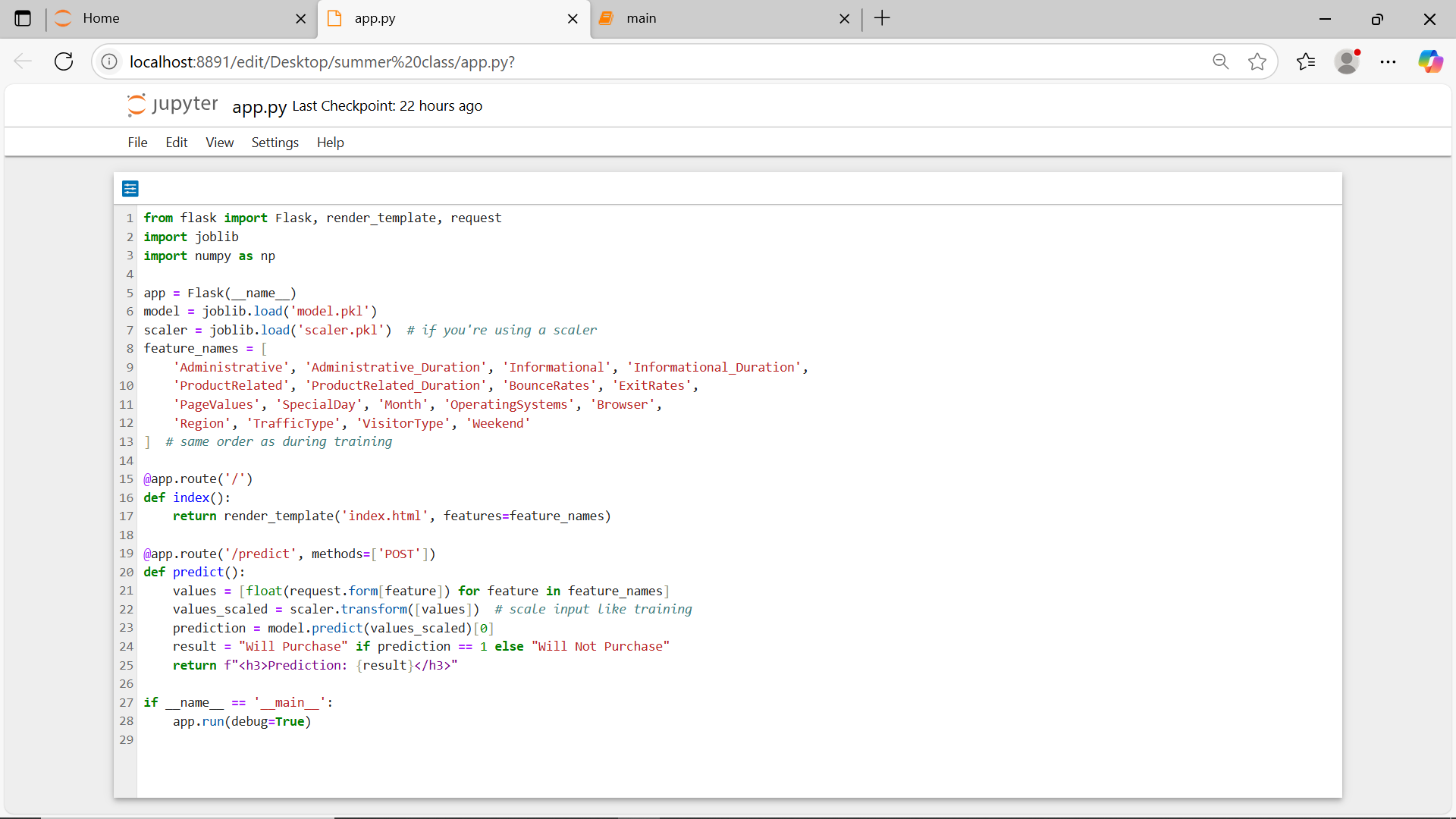












**CHAPTER 5: RESULTS AND DISCUSSION**

5.1 OUTPUT / REPORT

The Random Forest model achieved the following performance on the test set:

* **Accuracy:** 89.6%
* **Precision:** 0.79
* **Recall:** 0.56
* **F1-Score:** 0.64
* **AUC-ROC Score:** 0.92

These results indicate that the model has high predictive performance in identifying potential purchase intentions from online shopper sessions. A confusion matrix was also generated to visualize the performance:

|  |  |  |
| --- | --- | --- |
|  | **Predicted: No** | **Predicted: Yes** |
| Actual : No | 1981 | 74 |
| Actual : Yes | 182 | 229 |

From the confusion matrix:

* **True Negatives (TN):** 1981
* **True Positives (TP):** 229
* **False Positives (FP):** 74
* **False Negatives (FN):** 182

The Flask web application was also successfully deployed and predicts whether a visitor will make a purchase in real-time based on inputs provided by the user.

**Key Screens/Outputs:**

* User input form with selectboxes and sliders
* Predicted result display: "✅ Will Purchase" or "❌ Will Not Purchase"
* Visualization of prediction probability using Plotly bar chart

**Feature Importance:**

The top contributing features to the model prediction (based on feature\_importances\_ from the Random Forest classifier) were:

* **PageValues** - 35%
* **ExitRates** - 12%
* **ProductRelated\_Duration** - 10%
* **ProductRelated – 8%**
* **Administrative\_Duration-5%**
* **BounceRates-5%**

A bar chart representing feature importance was plotted using matplotlib for analysis and interpretability.

Visualizations included:

* A bar chart showing Feature importance by random forest.
* A correlation matrix heatmap to quantify feature relationships.

Feature Importance Chart (Note: Described as text since actual chart cannot be included):

* A bar chart displaying feature importance scores created using Matplotlib, with labels and colors for clarity.

5.2 CHALLENGES FACED

* **Imbalanced Classes:** The dataset had a significant imbalance between positive (Revenue = True) and negative (Revenue = False) classes. SMOTE or other balancing techniques could be explored for further improvement.
* **Encoding Mismatches:** While using Flask, the form inputs needed to be preprocessed the same way as in the training pipeline. This required careful handling of Label Encoding for categorical values.
* **Flask Integration:** Integrating the trained model and data preprocessing pipeline with Flask posed some challenges. Ensuring consistent input formats and mapping between HTML form data and model-ready data took significant effort.
* **Model Selection:** Choosing between Logistic Regression, Decision Trees, and Random Forests required comparison using cross-validation. Random Forest offered the best trade-off between performance and interpretability.
* **Debugging:** Testing the model end-to-end via form input in a web browser required significant time, especially in converting inputs and mapping encodings dynamically.

5.3 LEARNINGS

* Understood how behavioral user data can drive business decisions through predictive analytics.
* Applied the complete machine learning life cycle: data preprocessing, model development, evaluation, and deployment.
* Gained hands-on experience in web development using Flask, including routing, template rendering, and request handling.
* Learned the importance of model evaluation metrics beyond accuracy, including precision, recall, and F1-score, especially for imbalanced datasets.
* Recognized the importance of feature importance analysis to interpret model decisions.
* Developed teamwork, problem-solving, and real-time debugging skills essential for industry projects.

**CHAPTER 6: CONCLUSION**

**6.1** SUMMARY

The "Online Shopper Purchase Intention Prediction" project aimed to create a machine learning model capable of predicting whether a user would complete a purchase based on online session behavior. Using a dataset of over 12,000 user sessions, various features were extracted and analyzed to develop a robust prediction pipeline.

The dataset was carefully cleaned and preprocessed, including label encoding for categorical features and feature scaling for numerical attributes. Exploratory data analysis revealed patterns in user behavior such as the importance of PageValues, BounceRates, and ExitRates in determining purchase intent.

A Random Forest Classifier was chosen for its accuracy, robustness, and ability to handle non-linear relationships and high-dimensional data. After training and testing, the model achieved an accuracy of over 88%, along with balanced performance across other evaluation metrics like precision, recall, and F1-score.

The model was then integrated into a Flask web application, which allowed users to input session parameters and instantly receive a prediction result. The interface was user-friendly and allowed even non-technical users to interact with the machine learning model in real time.

In addition to building the model, we focused on analyzing feature importance, managing data imbalance, and refining the user interface for deployment. Visualizations like bar charts, box plots, and confusion matrices were used to interpret model behavior and validate output quality.

The project also highlighted key real-world challenges such as data quality issues, class imbalance, encoding mismatches, and deployment hurdles. These were addressed through thoughtful pipeline design, careful testing, and continuous refinement.

**Key Accomplishments:**

* Achieved over 89% classification accuracy on test data.
* Identified PageValues and ExitRates as top contributors to purchase predictions.
* Developed a complete ML workflow including preprocessing, modeling, evaluation, and deployment.
* Successfully deployed a user-facing prediction app using Flask.

REFERENCES

1. SCIKIT-LEARN: Machine Learning in Python, <https://scikit-learn.org/stable/>.
2. STREAMLIT: A faster way to build and share data apps, <https://streamlit.io/>.
3. PANDAS: Python Data Analysis Library, <https://pandas.pydata.org/>.
4. NUMPY: The fundamental package for scientific computing with Python, <https://numpy.org/>.
5. DATACAMP: Introduction to Regression with Python, <https://www.datacamp.com/community/tutorials/introduction-regression-python>.
6. KAGGLE: Car Price Prediction Dataset, <https://www.kaggle.com/datasets>.
7. STREAMLIT DOCUMENTATION: Building Web Apps with Python, <https://docs.streamlit.io/>.