FLiT

Data Project

Projects

Part A

Month 1 – Month 6

- **Project 1:** Market Basket Analysis for E-commerce
- **Project 2:** Hotel Reservation Analysis in SQL and Tableau.
- **Project 3:** Sentiment Analysis for Product Reviews.
- **Project 4:** Insurance Analysis with BigQuery, Tableau, and Google Data Studio.
- **Project 5:** Churn Prediction Model and Heroku Deployment.
- **Project 6:** Time Series Forecasting for Sales.

Part B

Month 7 – Month 12

- **Project 7:** Basic Fraud Detection with Supervised Learning.
- **Project 8:** Advanced Fraud Detection with Anomaly Detection
- **Project 9:** Implementing Cutting-Edge Fraud Detection with Deep Learning.
- **Project 10:** Enhancing Player Retention through A/B Testing in "Cookie Cats" Mobile Game.
- **Project 11:** Machine Learning for Loan Default Prediction with Taipy Integration.
- **Project 12:** Deep Learning for Advanced Loan Default Prediction with Taipy Integration.

Market Basket Analysis for E-commerce

Imagine you are working for a retail company, and you have access to a dataset containing customer transactions. Your task is to perform market basket analysis to uncover patterns in customer purchasing behavior. By identifying which products tend to be bought together, the company can make informed decisions to improve sales and customer satisfaction.

Objective

The goal of this project is to introduce you to the concept of market basket analysis, which is a crucial aspect of data science in retail and e-commerce. You will learn how to extract valuable insights from transaction data, understand customer purchasing behaviour, and use this knowledge for business optimization.

Tools

- Data Analysis Tool: Python (using libraries like Pandas, etc.)
- Data Visualization Tool: Matplotlib, Seaborn
- Scikit-Learn
- Jupyter Notebook

High-Level Steps:

1. Data Preparation
2. Exploratory Data Analysis (EDA)
3. Market Basket Analysis
4. Visualization
5. Interpretation and Insights
6. Recommendations
7. Presentation

Deliverables:

- A well-documented Jupyter Notebook or report containing code and explanations.
- Visualizations that support your findings.
- A presentation or report for your mentorship group.
- A GitHub repository with documentation and code.

Click for dataset

Hotel Reservation Analysis in SQL and Tableau

Objective

In this open-ended project, you'll work with a hotel reservation dataset to gain insights and create visualizations using your choice of SQL environment for data manipulation and Tableau for visualization. The goal is to explore the dataset, identify interesting patterns, and create impactful visualizations to tell a compelling data story.

Tools

- SQL
- Tableau
- Github

Your task is to work with a hotel reservation dataset that contains information about reservations at two types of hotels: Resort Hotels (H1) and City Hotels (H2). You have the flexibility to choose your preferred **SQL** environment (e.g., MySQL, PostgreSQL, BigQuery, Athena) for data manipulation and **Tableau** for visualization.

High-Level Steps:

1. Data Import
2. Data Exploration
3. SQL Analysis
4. Tableau Visualization
5. GitHub Repository
6. Tableau Dashboards publishing
7. Presentation

Deliverables:

- SQL environment with imported dataset and SQL scripts.
- Published tableau dashboard.
- A presentation or report summarizing your findings and recommendations.
- A GitHub repository with documentation and code.

Click for dataset

Sentiment Analysis for Product Reviews

Objective

In this project, you will delve into the world of natural language processing and sentiment analysis. You will work with a dataset of product reviews to analyze and classify the sentiment of each review as positive, negative, or neutral. This project aims to enhance your understanding of text analysis and machine learning for sentiment classification.

Tools

- Python libraries: Pandas, etc.
- Machine Learning Library: Scikit-learn, Spacy, etc.
- Jupyter Notebook for documentation
- GitHub
- Streamlit

As an analyst, you will work with a dataset of product reviews from an e-commerce platform. The dataset includes text reviews and associated ratings, which you will use to perform sentiment analysis. The goal is to build a model that can classify each review as positive, negative, or neutral based on the text content.

High-Level Steps:

1. Data Preparation
2. Exploratory Data Analysis (EDA)
3. Data Preprocessing
4. Sentiment Labeling
5. Text Vectorization
6. Model Building
7. Model Evaluation
8. Sentiment Analysis Dashboard

Click for dataset

Deliverables:

- SQL environment with imported dataset and SQL scripts.
- Published tableau dashboard.
- A presentation or report summarizing your findings and recommendations.
- A GitHub repository with documentation and code.

Car Insurance Analysis with BigQuery, Tableau, and Google Data Studio

As an analyst in a Car Insurance company, your role will be to analyze this data and craft insightful dashboards designed for insurance analysis. We'll cater to two distinct audiences: upper management, who will use Tableau, and operational teams, who will rely on Google Data Studio.

Objective

In this project, you will act as insurance analysts. You will import a hotel reservation dataset into Google BigQuery, create dashboards in Tableau for upper management, and another operational dashboard in Google Data Studio. The project aims to showcase data analysis, visualization, and data pipeline skills.

Tools

- Google BigQuery
- Tableau
- Google DataStudio
- GitHub

High-Level Steps:

1. Setting Up Data in Google BigQuery
2. Connect BigQuery to Tableau
3. Create important metrics for Management reporting
4. Connect BigQuery to Tableau
5. Tableau Dashboards publishing
6. Presentation

Deliverables:

- Google BigQuery dataset with the car insurance data.
- A Tableau dashboard tailored for upper management.
- A Google Data Studio dashboard designed for operational use.
- A GitHub repository with documentation and code.

Click for dataset

Churn Prediction Model and Heroku Deployment

As data scientist for a telecoms company, you will work with a telecom company's customer data, including features like gender, tenure, internet service, payment method, and churn status. Your goal is to create a model that predicts whether a customer is likely to churn (leave the service). Additionally, you will deploy this model on Heroku, allowing for real-time predictions.

Objective

In this project, you will build a machine learning model to predict customer churn for a telecom company. The project includes data preprocessing, model training, and deploying the model on Heroku, making it accessible online.

Tools

- Python libraries: Pandas, Scikit-Learn, etc.
- Heroku
- GitHub

High-Level Steps:

1. Data Exploration and Preprocessing
2. Machine Learning Model
3. Heroku Deployment
4. User Interface
5. Tableau Dashboards publishing
6. Documentation and Presentation

Deliverables:

- Jupyter Notebook or script containing data preprocessing, model training, and evaluation.
- Deployed churn prediction model on Heroku.
- Documentation explaining the project and how to use the deployed model.
- A web-based user interface for interacting with the model.

Click for dataset

API-Driven Historical Weather Data Analysis with Weatherstack and Tableau

Objective

In this project, You will harness the Weatherstack API to collect historical weather data for locations around the world. You will build a data pipeline for retrieving weather data, store it in a PostgreSQL database, and leverage Tableau for insightful data visualizations.

Tools

- Python
- API – Weatherstack
- PostgreSQL
- GitHub
- Tableau

You will put the Weatherstack API to work to gather historical weather data for various locations. This data includes essential factors like temperature, precipitation, humidity, and more. You will establish a data processing pipeline that fetches this information, stores it in a PostgreSQL database, and conducts a comprehensive historical weather analysis. Furthermore, You will use Tableau for creating engaging visualizations that bring the data to life.

High-Level Steps:

1. Data Collection via Weatherstack API
2. Data Storage
3. Data Ingestion
4. Historical Weather Analysis
5. Tableau Data Visualization
6. Documentation and Presentation

Deliverables:

- A Python script for interfacing with the Weatherstack API to retrieve historical weather data.
- A PostgreSQL database containing the gathered historical weather data.
- SQL queries or analysis scripts for conducting the historical weather analysis.
- Engaging Tableau visualizations that effectively convey the outcomes of the analysis.
- Comprehensive project documentation and a presentation to communicate your findings and insights.

Basic Fraud Detection with Supervised Learning

Imagine you've just joined a financial institution as a data scientist. The company faces increasing challenges with credit card fraud, and there's a pressing need for a fraud detection system. In this initial project, your primary objective is to establish a baseline fraud detection system. This system will lay the foundation for more advanced techniques in subsequent projects and provide a starting point for identifying potentially fraudulent transactions.

Objective

These projects offer a pragmatic journey from foundational fraud detection to the implementation of advanced deep learning techniques.

Tools

- Python libraries e.g. scikit-learn, etc
- Flask
- GitHub

High-Level Steps:

1. Data Collection
2. Data Preprocessing
3. Model Selection
4. Model Training
5. Evaluation Metrics
6. Deployment

Deliverables:

- A Flask app hosting the basic fraud detection model.
- Documentation detailing the data preprocessing steps, model selection, training, and performance evaluation.

Click for dataset

Advanced Fraud Detection with Anomaly Detection

As the institution's financial data scientist, you've successfully deployed the baseline fraud detection system, but it's time to make it even more robust. Your objective in this intermediate project is to enhance the sensitivity and accuracy of fraud detection. You'll achieve this by incorporating anomaly detection techniques. These methods are particularly valuable for identifying rare and novel fraud patterns that might have escaped traditional approaches.

Objective

These projects offer a pragmatic journey from foundational fraud detection to the implementation of advanced deep learning techniques.

Tools

- Python libraries e.g., scikit-learn, etc
- Flask
- GitHub

High-Level Steps:

1. Data Preprocessing
2. Model Selection
3. Model Training
4. Ensemble Learning
5. Evaluation Metrics
6. Deployment

Deliverables:

- A Flask app hosting the advanced fraud detection model combining supervised and anomaly detection techniques.
- Documentation detailing the data preprocessing enhancements, model selection, training, ensemble learning, and evaluation.
- A README on the model's improved performance and its effectiveness in identifying emerging fraud patterns.

Click for dataset

Implementing Cutting-Edge Fraud Detection with Deep Learning

Objective

These projects offer a pragmatic journey from foundational fraud detection to the implementation of advanced deep learning techniques.

Tools

- Python libraries e.g., scikit-learn, etc
- Flask
- GitHub

In this advanced project, you'll elevate the institution's fraud detection system to the next level. Your objective is to explore deep learning techniques, specifically neural networks, to build a highly sophisticated real-time fraud detection system. This deep learning approach allows for the creation of a system with improved accuracy, capable of detecting even the most subtle and evolving fraud patterns.

High-Level Steps:

1. Data Preprocessing
2. Model Selection
3. Model Training
4. Hyperparameter Tuning
5. Evaluation Metrics
6. Deployment

Deliverables:

- A Flask app hosting the deep learning-based real-time fraud detection system.
- Comprehensive documentation detailing data preprocessing for deep learning, model selection, training, hyperparameter tuning, and evaluation.
- A README showcasing the deep learning model's performance improvements and its ability to detect subtle and emerging fraud patterns, offering the institution a cutting-edge fraud detection solution.

Click for dataset

Enhancing Player Retention through A/B Testing in "Cookie Cats" Mobile Game

Objective

In this project, we will be leveraging A/B Testing for mobile Game Optimization

Tools

- Python libraries e.g., Pandas, etc
- Flask
- GitHub

Imagine you're working for a mobile game development company, and one of your most popular games, "Cookie Cats," is experiencing a drop in player retention. To address this issue, you aim to employ A/B testing to experiment with different game design changes. The primary objective of this project is to determine which game design modification results in better player retention.

High-Level Steps:

1. Data collection
2. Data Exploration
3. A/B Test Setup
4. Data Analysis
5. Statistical Analysis
6. Results Interpretation
7. Recommendations

Deliverables:

- Jupyter Notebook or script containing A/B testing.
- A comprehensive README detailing the A/B testing process, results, and recommendations

Click for dataset

Machine Learning for Loan Default Prediction with Taipy Integration

Objective

In this project, we will be Developing an Effective Machine Learning Model to Predict Loan Defaults and Deploying on Taipy

Tools

- Python libraries e.g., Scikit-Learn, etc
- Taipy
- GitHub

In this initial AI project, your mission is to assist a financial institution in addressing the critical challenge of loan defaults. The objective is to create a robust machine learning-based system that can predict the likelihood of loan defaults, ultimately enhancing credit risk management. This project marks the first step toward more advanced AI-driven solutions.

High-Level Steps:

1. Data collection
2. Data Preprocessing
3. Feature engineering
4. Model Selection
5. Model Training
6. Hyperparameter Tuning
7. Evaluation Metrics
8. Model Deployment on Taipy

Deliverables:

- A real-time loan default prediction system deployed on Taipy.
- A comprehensive README on data preprocessing, feature engineering, model selection, and evaluation.

Click for dataset

Deep Learning for Advanced Loan Default Prediction with Taipy Integration

Objective

In this project, we will be Elevating Loan Default Prediction through Deep Learning Techniques and Deploying on Taipy

Tools

- Python libraries e.g., Scikit-Learn, etc
- Taipy
- GitHub

Building on the success of the previous project, your focus shifts to deep learning techniques. The goal is to harness deep learning for loan default prediction, capturing complex patterns within the data to further enhance predictive accuracy. This advanced project involves deploying the deep learning model on the Taipy platform for real-time predictions.

High-Level Steps:

1. Data Preprocessing
2. Model Selection
3. Model Training
4. Hyperparameter Tuning
5. Evaluation Metrics
6. Model Deployment on Taipy

Deliverables:

- An advanced real-time loan default prediction system deployed on Taipy..
- A comprehensive README on data preprocessing, feature engineering, model selection, and evaluation.

Click for dataset