Page Summary

Executive Summary

- The dashboard shows the summary of Power BI developer salary from 10 different countries
- The data is represented in the Pie Chart and Bar Graph

Projects

Six machine learning models have been developed, including **supervised classification** and **supervised regression models**

1. Supervised Classification Model - SDA Face Recognition

The primary goal of this project is to develop a robust, easy-to-use face recognition system capable of identifying individuals from a small, predefined set of categories. This system is intended to be used in various scenarios, such as security systems, personalized access control, and automated identity verification processes. The model's ability to classify images in real-time allows for immediate feedback and efficient user interaction.

The project employs the following technologies:

1. **TensorFlow and Keras**: For developing and deploying the deep learning model.
2. **Streamlit**: For building an interactive web application that allows users to upload images and receive instant predictions.
3. **Python Libraries**: Including **NumPy, PIL** and **TensorFlow** utilities for image processing and model deployment.

Application to CGC: Customer verification

Purpose: Use face recognition as an added layer of security for authenticating customers who interact with CGC services (e.g., online loan applications, claims, etc.).

How: When customers submit their identity information (such as during loan application submissions), we could have them upload a photo. The model will verify the individual by matching the uploaded photo with a stored database of authorized personnel or customers. This helps reduce fraud and ensures that the person applying for a service is who they claim to be.

2. Supervised Classification Model - Phishing Email Detection

The main goal of this project is to develop a robust and efficient solution for detecting phishing emails from a large dataset of text-based messages. Phishing emails, which are often used for fraudulent activities like stealing personal information or spreading malware, are a common threat. This system is intended to automate the detection process, enabling users to quickly and accurately identify phishing attempts with minimal human intervention.

The project employs several technologies and tools to build and deploy the model:

1. **Python Libraries**:
    - **Pandas**: For data manipulation and handling the dataset.
    - **NumPy**: For numerical operations.
    - **Scikit-Learn**: For machine learning, including model training, feature extraction (**CountVectorizer**), and classification (**Multinomial Naive Bayes**).
2. **Streamlit**: For building an intuitive web interface, allowing users to interact with the model and receive predictions in real-time.
3. **CountVectorizer**: Used for converting text data (email messages) into numerical features suitable for the machine learning model.

Application to CGC: Email security for internal and external Communications

Purpose: Protect employees from phishing attacks in emails, especially since CGC likely deals with sensitive financial and personal information.

How: How: The model can be used to automatically analyze incoming emails to CGC employees. Emails flagged as "Phishing" can be routed to a security team for further analysis or can trigger an automatic alert to warn the recipient. This could protects CGC employees from opening fraudulent emails that could lead to data breaches, financial loss, or malware installation.

3. Supervised Regression Model - Salary Prediction Tool

- This tool can be utilized to assist user to gauge the salary of a Power BI developer from different country
- User needs to specify the country, qualification and years of working experience
- The accuracy of this prediction tool is roughly 95%

The project employs several technologies to build and deploy the model:

1. **Streamlit**: Utilized to build an interactive web app that allows users to input their details and get salary predictions in real time.
2. **Pickle**: Used to load a pre-trained machine learning model (regressor), as well as the label encoders (le_country and le_education) that map categorical variables (country and education level) into numerical values required by the model.

3. **NumPy**: Used for array manipulation and ensuring that the input data is in the right format (array) for the model to process.
4. **Machine Learning Model (Regressor)**: The regressor model is a machine learning model that has been trained on historical salary data and is used to predict the annual salary based on user inputs.

Application to CGC: Salary benchmarking for Power BI developers

Purpose: With CGC being a major financial institution, it is essential to have competitive and accurate salary offerings for tech roles. The model can be used to benchmark the salary expectations for Power BI developers based on factors like **experience**, **country**, and **education**.

How: HR departments could use this tool to estimate appropriate salary ranges when hiring or negotiating salaries for developers, ensuring CGC stays competitive in the job market.

4. Supervised Classification Model - Loan Default Prediction

The goal of this ML model is to predict whether a loan applicant will default on their loan (`LoanDefault = 1`) or not (`LoanDefault = 0`) based on various business and financial features. The prediction is based on patterns learned from historical data. This prediction can help financial institutions assess the risk of lending and make informed decisions.

There are 3 algorithms compared in term of their performance and the best algorithm is selected to predict loan default based on the AUC-ROC score:

1. **Random Forest Classifier**: A robust ensemble learning method that combines multiple decision trees to improve prediction accuracy.
2. **Gradient Boosting Classifier**: An ensemble method that builds decision trees sequentially, where each tree corrects the errors of the previous one.
3. **XGBoost Classifier**: An optimized implementation of gradient boosting that is faster and more efficient.

Application to CGC: On-the-spot loan risk assessment

Purpose: Branch officers can use this tool directly during loan application processing to predict the likelihood of a loan default based on the applicant's business data.

How: When a small business approaches the branch for a loan guarantee, the loan officer can enter details (like business age, revenue, loan amount requested, etc.) into the model. The model will then predict the default risk, helping the officer to decide if the loan application should proceed. This can significantly speed up loan processing for faster decision-making and reduced back-and-forth communication between branches and headquarters.

5. Supervised Regression Model - Loan Eligibility Prediction

- A Loan Eligibility Prediction model using a **Random Forest Classifier** was developed to automate and enhance the loan approval process.
- The model uses key features like **income, credit score, and loan amount** to predict eligibility and trained on a **synthetic dataset of 50,000 samples**
- The benefit of this model is **reduces human bias, improves decision speed**, and **minimizes loan default risks**, making the approval process more efficient and reliable.

- The impact of this model is supports multiple departments i.e. **Credit Risk, Loan Processing, Customer Service, Marketing, and Finance**—with faster, data-driven insights and improved workflow.

The project employs several technologies to build and deploy the model:

1. **Streamlit**: Utilized to create a web interface for easy user interaction, allowing them to input business details and receive loan eligibility predictions.
2. **Joblib**: Used to load pre-trained machine learning models (loan_eligibility_model.pkl) and the corresponding feature list (model_features.pkl) for prediction.
3. **Pandas**: Used to organize and structure the user inputs into a DataFrame format compatible with the model's expected input.
4. **Machine Learning Model**: A classifier model trained to predict loan eligibility based on business attributes such as credit score, annual revenue, and industry.

Application to CGC: Streamlining loan guarantee decisions

Purpose: CGC's loan officers can use this app at the branch level to quickly assess which businesses meet the eligibility criteria for loan guarantees.

How: By entering business data, the branch staff can instantly know if the applicant qualifies for CGC's loan guarantee and speeding up the application process.

6. Supervised Regression Model - SME Loan Eligibility Amount Predictor

- Developed a machine learning model (SME Loan Eligibility Amount Predictor) to estimate eligible loan amounts based on business-specific attributes like company size, business age, and employee count.
- Utilized a Random Forest Regressor trained on a 50,000-sample dummy dataset generated by CoPilot.
- Designed to support Branch Sales Management, Credit Risk, and Product Development teams in financial institutions.
- Aims to streamline SME loan evaluations, enhance credit decision-making, and improve customer accessibility.

The project employs several technologies to build and deploy the model:

1. **Streamlit**: Utilized for creating an interactive web interface for users to input data and view the results.
2. **Joblib**: Used to load the pre-trained machine learning model (sme_loan_model.pkl) and the LabelEncoder objects for encoding categorical features.
3. **Pandas**: Used to manage and structure input data into a format compatible with the model.
4. **Scikit-learn**: Utilized for encoding categorical features (LabelEncoder) and for prediction tasks with the pre-trained model.

Application to CGC: Loan eligibility prediction for SMEs

Purpose: Streamline the loan application process and help SMEs understand their eligibility before applying.

How: Allow SMEs to enter their details into the app, which will predict their eligibility for loan guarantees, reducing manual work for CGC staffs.