Gulf Bank Datathon 2022

Data Modeling Track Challenges

Setup:

- You will be using a retail banking dataset found on data.world, download it from here and set up your project.

Transactions Path:

Level 1: Exploring (1 point per task)

1. How many transactions are there in the dataset?
2. What is the value of the largest transaction made and to which account/accounts was it made?
    a. Get the name, gender, age, and the city this client lives in
3. Which account(s) had the largest balance and what date was this? How does this compare to the balance of this account to the latest date?
4. Create a histogram and box plot displaying the amounts distribution
5. Split the amounts into bins:
    a. '<100'
    b. '100-250'
    c. '250-500'
    d. '500-1K'
    e. '1K-2K'
    f. '2K-4K'
    g. '4K-8K'
    h. '8K-15K'
    i. '15K-30K'
    j. '30K-60K'
    k. '>60K'
6. How many debit and credit transactions are there?
7. What are the different types of transaction operations?
8. What are the different K Symbol values?
9. Which top/bottom 3 banks have the highest count of transactions?
10. Which top/bottom 3 banks have the highest value of transactions?

Level 2: Analyzing (3 points per task)

1. Which Month has the most transactions? Did this differ over the years? Visualize this.
2. Which day of the month has the most transactions? Did this differ over the years? Visualize this.
3. Which day of the week has the most transactions? Did this differ over the years? Visualize this.

4. Is there a trend or shift in type of transactions over time? Visualize the count & sum of different type of transactions over time.

Level 3: Modeling

1. Create a model to predict the transaction operation **(15 points)**
   a. Try out at least 3 different models, and comment on why you chose these models
   b. Measure accuracy (at least one model should have accuracy of 80% or higher)
2. Time to clean up your work. Organize your notebook, gather your findings, prepare recommendations and get ready to present them. **(5 points)**
3. **Bonus:** Repeat the same process by creating a model to predict the transaction "k_symbol"

Loans Path:

Level 1: Exploring (1 point per task)

1. What are the different durations that loans are taken in and how many loans are taken for each duration?
2. What are the different statuses for the loans?
3. What are the different loan purposes and what is the most common purpose?
4. What is the median loan amount?
5. How is the median loan amount changing over the years? Visualize this
6. What are the top 3 most common locations that people take loans in? (The location is represented by a number)
7. What is the average value of loans for each duration? Visualize this.

Level 2: Analyzing (3 points per task)

1. How many of the loans taken are by males and how many of them are by females?
2. Create categories for the ages of the clients
   a. <25
   b. 25-35
   c. 35-45
   d. 50-65
   e. >65
3. Check which age group has the highest count of loans? Visualize this
4. Check which age group has the highest value of loans? Visualize this
5. Create categories for the loan "amounts":
   a. '-75K'
   b. '75K-150K'

  c. '150K-250K'

  d. '250K-350K'

  e. '350K+'

6. What are the different frequencies for loan payments?

Level 3: Modeling

1. Create a model to predict the transaction purpose **(10 points)**
    a. Try out at least 3 different models, and comment on why you chose these models
    b. Measure accuracy (at least one model should have accuracy of 80% or higher)
2. Create a model to predict the transaction "amount category" **(10 points)**
    a. Try out at least 3 different models, and comment on why you chose these models
    b. Measure accuracy (at least one model should have accuracy of 80% or higher)
3. Time to clean up your work. Organize your notebook, gather your findings, prepare recommendations and get ready to present them. **(5 points)**

Luxury Loan Portfolio Path:

Level 1: Exploring (1 point per task)

1. What are the min, max, and median funded amounts?
2. What are the min, max, and average interest rates?
3. What are the min, max, and median, average property value?
4. What are the different "durations months" for the loans and how many loans are under each one?
5. What are the different purposes for the luxury loans and how many loans are there per purpose?
6. What are the 3 most common "total units" found in a property?
7. What are the 3 most common "building class categories" loans are taken for?

Level 2: Analyzing (3 points per task)

1. Create bins for property value:
    a. '-1M'
    b. '1M-2M'
    c. '2M-3M'
    d. '3M+'
2. Create bins for employment length:
    a. '-5'
    b. '5-10'

    c. '10-15'

    d. '15+'

3. Analyze how interest rates differ according to different loan purposes. Visualize this.
4. Identify the building class category that is most profitable.
    a. Using funded amount and interest rate
5. Which "purpose" generates the most value?

Level 3: Modeling

1. Create a model to predict the "property value" **(15 points)**
    a. Try out at least 3 different models, and comment on why you chose these models
    b. Measure accuracy (at least one model should have accuracy of 80% or higher)
2. Create a model to predict the transaction "employment length" **(5 points)**
    a. Try out at least 3 different models, and comment on why you chose these models
    b. Measure accuracy (at least one model should have accuracy of 80% or higher)
3. Time to clean up your work. Organize your notebook, gather your findings, prepare recommendations and get ready to present them.