**OBJECTIVE FOR DOING PERSONAL PROJECTS**

* Challenge myself to work outside my comfort zone – learn new concepts beyond my everyday work and apply.
* Challenge myself to plan a project, execute it and pitch it within a specified timeline. Help me overcome procrastination and maintain a routine in developing my skills in my career. Beyond reading - Practice

**SOFTWARE DEVELOPMENT SKILLS FOR DATA SCIENTISTS**

* Writing clean, modular and reusable code
* Documentation
* Version Controlling
* Testing
* Logging

**Situation**

**Task**

**Action**

Business Context Understanding

* Customer user journey
* Request business headlines from client
  + Time period for analysis
  + Total unique number of customers overall vs customer type
  + Total unique number of sales agents
  + Total unique number of products
  + Total revenue/sales amount
  + Average revenue/sales amount

Business Problem Understanding

* Project statement/definition
* Project requirements/objectives
* Architectural specification
  + For a machine learning model solution:
    - User interface
    - Data flow through program
  + For actionable insights solution:
    - The use case for actionable insights generated eg: investor pitch, decision making
    - The stakeholders/ audience
    - The dimension/ verticals in the data to emphasize or highlight in the narrative

Data Understanding

* Did a sanity check of the customer and transactions data I was provided:
  + - first and last dates,
    - dates consistency,
    - date occurrence sequence,
    - total number of records,
    - total number of features,
    - total number of unique customers, missing values, outliers in context of business. Build context of business (eg products, business headlines etc) using historical data to confirm business headlines communicated by client

Built pipeline and iterated:

* Explore data focusing on verticals of interest
* Ask Key Performance Questions and note possible metrics/KPIs of interest
* Clean data
* Derive additional features
* Slice data along verticals of interest
* Determine best way to visualize and present insights

**Results**

**Impact**

**2. Value and things learnt/ Techniques Studied**

* **Data cleaning**
* **Webscraping for external data (zipcode for length of mile)**
* **365 data science: customer analytics + time series**
* **Know metrics series**
* **Slides**
* **Market basket analysis**
* **Python + Dask**
* **SQLAlchemy**
* **Business Analytics**
* **Productionize (flask wrapper + make it an API endpoint integrated into my website)**
* **Ensemble modeling**

**3. [[[[[ANOTHER PROJECT FOCUSED ON CREDIT SCORING USING SCIKIT LEARN AND PYCARET]]]]]**

**4. [[[[[ANOTHER PROJECT USING GRAPHS AND NETWORK ANALYSIS]]]] – my interaction on social media or whatsapp page**

**5. Data product for NGO**

**6. Collect personal data and do analysis on it**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Why did I choose to do project about this? |  |  |  |  |
| What does this project mean to me? |  |  |  |  |
| What was your favourite thing about working on this project? |  |  |  |  |
| What was your least favourite thing about working on this project? |  |  |  |  |
| What technical challenges dis you face during this project and how did you overcome them? |  |  |  |  |
| Where did you get this data set and what techniques did you use to clean the data? |  |  |  |  |
| Why did you choose to use the statistical technique you used for this project? |  |  |  |  |
| Could you explain how this algorithm/statistical technique/ section of code works? |  |  |  |  |
| What libraries, packages or other tools did you use for this project? |  |  |  |  |
| How long did it take you to put this project together? |  |  |  |  |
| If asked to, how might you expand on this project? |  |  |  |  |
| If you had to do it again, what might you change about this project? |  |  |  |  |
| How will the skills you used on this project be valuable to our business? |  |  |  |  |
| (If group project) What was your job on this project? |  |  |  |  |
| (If group project ) How was this project orgarnized and version-controlled? |  |  |  |  |
| (If group project) Can you talk about a conflict or disagreement you had with teammates during this project and you overcame it? |  |  |  |  |

Data Cleaning Project

#Data Storytelling Project

Switching between tools to get the work done fastests

== ASSUMPTION ABOUT Me:

#Role

I am an expert analyst who has been hired to discover with speed and accuracy, interesting and important opportunities worth exploring

further and business problems worth solving from the business data

#Steps

1. Read up on business and industry modus operandi

2. Understand the data (data dictionary not provided)

3. Clean the data

4. Jot down interesting initial insights from the data (creating CONTENT)

5. Determine focus metric and create story that resonates with this focus metric (sales and employee performance)

6. Draw a dashboard of what I expect to create with a visualizing tool (upload snapshot to Github account) (deciding on the VISUALS)

7. Create a dashboard with Tableau Public and create a Presentation Deck with the snapshots from the Tableau Dashboard. Some features may be

derived in Tableau. Prefix them with derived.

Goal is to craft a compelling data story or narrative that will inpire decision-makers to take an action ie invest in

exploring the opportunities and business problems I recommend in my presentation. Craft your data story to be laser focused on

showing and communicating your recommendations (focus metric) so drop all the not-so strong points from the visual display

(DELIVERY)

#SUMMARY of 3 to 6:

a. What is the expectation of the client? That becomes the pivot or focus of discovering when exploring the data

b. What does the past data say about this expectation / focus comparing it across branches, time etc. Historically what has been the trend.

c. What are the statistical projections regarding this expectations. The header should state the recommendation and the visuals should

support the recommendation

2018.12.01\_Yvette\_Data Visualization and Storytelling (external hard drive)

#Ideas

A

- Data Extraction Date

- Timeframe

- Users (Customers) = X+ (discount, loyal, need-based customers)

- Distinct Items = 452

- Distint Categories = 20

- Revenue

B

- Inventory dashboard (modify date indicates date products were received from the manufacturer distributors)

- Performance dashboard - Sales, Employees

- Growth dashboard (customers, sales, revenue, employees, platform engagement etc)

- Usage dashboard

- <https://blog.close.com/sales-analysis/>

- <https://towardsdatascience.com/data-driven-growth-with-python-part-1-know-your-metrics-812781e66a5b>

C

- Decide on metric for quantifying business problem

- Give topline numbers

- Slice topline numbers according to region, customer type, sales day of week, sales week of month, sales time of day

#Future Work

1. Convert data cleaning notebook to general script that can scale in cleaning similar data

2. Clean pre\_process.py

3. Customer Analytics

4. Time Series Analysis (Sales Forecasting)

5. Market Basket Analysis

Grocery Sales Data

Data Source

<https://www.kaggle.com/codemysteries/salesdb>

Clean Data Profile

* 98,759 (98k +) unique customers residing in 96 different cities. How many states?
* 23 unique employees residing in 18 different cities (what is the most occurring city? May indicate headquarters of business) and state
* Employees were hired between February 2010 and February 2017
* All the customers and employees were within the US
* 452 unique products grouped into 20 unique categories
* 6,758,125 (6M +) sales transactions
* Sales transactions were from 1st January 2018 on 9th May 2018 (4 months)

Data Audit and Data Cleaning Treatments

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Set | Feature | Data Quality Issue | Missing Data | | Treatment |
| Count | Proportion |
| Customer | MiddleInitial | Missing data | 977 | 0.01 | - |
| Employee | BirthDate | Non-standard format |  |  | Converted to standard datetime object |
|  | HireDate | Non-standard format |  |  | Converted to standard datetime object |
| Countries | CountryCode | Missing data | 1 | 0.00 | Imputed with ‘AV’ |
| Products | Resistant | Missing data | 140 | 0.31 | Replace with ‘Missing’ |
|  | IsAllergic | Missing data | 130 | 0.29 | Replace with ‘Missing’ |
|  | VitalityDays, | Spelt wrongly, missing data, | 274 | 0.61 | Renamed feature, replaced missing values with 0 |
|  | ProductName | Double spaces |  |  | Stripped double spaces |
|  | Price | Non-standard format |  |  | Generated new price list per category |
| Category | CategoryName | Categories were not exhaustive |  |  | Created new categories to fully categorize all products |
| Sales | Discount | Missing data | 5,406,931 | 0.80 | Replaced missing values with 0 |
|  | SalesDate | Non-standard format, missing data | 67,526 | 0.10 | Converted to standard datetime object, Filled missing data with backpropagation and sorting the dates |
|  | SalesID |  |  |  | Renumbered SalesID to reflect sequential ordering of transactions |
|  | TotalPrice | Non-standard format |  |  | Converted values to numeric values |

Feature Engineering

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data Set | Raw Feature | Data Type | Derived Feature | Definition/ Description |
|
| Employee | BirthDate | Datetime | derived.biological\_age |  |
|  |  | Datetime | derived.biological\_age.group |  |
|  | HireDate | Datetime | derived.service\_age.yrs |  |
|  |  | Datetime | derived.service\_age.yrs.group |  |
| City | CityName | String | derived.state |  |
|  |  | String | derived.region |  |
| Sales, Product | Quantity, Price, CategoryName | Integer, Numeric, String | derived.total\_spend.alltime |  |
| Sales | SalesDate | Datetime | derived.salesdate\_day\_of\_week |  |
|  |  |  | derived.salesdate\_time of\_day |  |
|  |  |  | derived.salesdate\_week\_of\_month |  |
|  |  |  | derived.customer\_type.frequency | first-time, repeat |
|  |  |  | derived.customer\_type.value | Loyal, discount, churning, high-value, low-value (since we don’t have date of registration we can’t be too quick to judge with this) |
|  |  |  | derived.unique\_visits.count.alltime |  |
|  |  |  | derived.total\_products\_bought.count.alltime | sum(quantity of products per visit) |
|  |  |  | derived.average\_products\_bought.count.alltime | derived.total\_products\_bought.count.alltime /derived.unique\_visits.count |
|  |  |  | derived.category\_with\_most\_spend.amount.alltime |  |
|  |  |  | derived.average\_spend.amount.alltime | derived.total\_spend.alltime /derived.unique\_visits.count |
|  |  |  | derived.lowest\_one-time\_spend.amount.alltime | Min(total spend per visit) |
|  |  |  | derived.highest\_one\_time\_spend.amount.alltime | Max(total spend per visit) |
|  |  |  | derived.standard\_deviation\_between\_visits.alltime | Std(between visits) |
|  |  |  | derived.standard\_deviation\_between\_spends.amount.alltime | Std(between all spends per visit) |
|  |  |  | derived.total\_discounts.count.alltime |  |
|  |  |  | derived.total\_discounts.amount.alltime |  |

* Is Class in Products data correlated with Quantity in Sales data?
* Get metrics for online commerce?
* Check category of products with VitalityDays, Class, Resistant and IsAllergic?

Business Topline Metrics

Customers

* Total number of customers/ users
  + First-time vs Repeat customers
  + Customer growth trend
* Demographic of customers

Employees

* Total number of employees
  + Employee growth trend
* Demographic of employees
* Gender distribution of employees
* Biological age distribution
* Service age distribution

Products

* Number of categories
* Average category unit price

Sales

* Total revenue (volume)
  + Total revenue growth trend month-on-month (volume)
  + Average revenue month-on-month
* Total unit of items bought per category
* Day of week, Time of day, Week of month with most frequent:
  + Sales/Revenue (volume and value)
  + Discounts (volume and value)

Customer Usage Metrics

* Average order size month-on-month
* Most popular category patronized per month
* Average order spend month-on-month
* Total spend month-on-month overlaid with average order size all time
* Average discount received month-on-month

Employee Performance Metrics

* Total sales per employee
* Average sales per employee per month
* Employee with most customer spend each month

KPQs and KPIs

Metrics

Building pipeline from start to finish and iterating