Capstone Project

Harnessing NLP to Detect Stress in Social Media

Early Intervention for Mental Wellbeing

Presenter: Jimmy Chong

© 2024 Institute of Data

Agenda

- Bio
- Project Context & Business Problem
- Business & Data Science Considerations
- Data Overview
- Data Exploration
- Data Split
- Data Overview
- Deliver
- Summary, Conclusions & Next Steps
- Appendix

© 2024 Institute of Data

Bio

- Education
- Professional experience
- Data science learnings and experience
- Relevance to the project

© 2024 Institute of Data

Project Context & Business Problem

- **Industry:** NLP and Mental Health
- **Problem:** Detecting stress in social media text
- **Interest:** Growing mental health concerns, potential for early intervention, NLP advancements
- **Previous Work:**
  - **NLP techniques:** Sentiment analysis, topic modeling, machine learning
  - **Domains:** Twitter, Reddit, general datasets
  - **Key findings:** Promising accuracy, varying generalizability and robustness
  - **Limitations:** Reliance on labeled datasets, potential for bias
  - **Contribution:** Exploring LSTM networks, addressing limitations

© 2024 Institute of Data

Business & Data Science Considerations

- **Stakeholders:** Mental health professionals, social media platforms, technology companies, individuals
- **Business Question:** Can we accurately detect stress in social media text?
- **Business Value:** Early intervention, improved user experience, new market opportunities
- **Data:**
  - **Question:** Can we extract effective features for stress prediction?
  - **Required:** Social media text, stress labels, metadata
  - **Sourced:** Reddit, Twitter, Kaggle
  - **Generation:** User-generated content
  - **Future Sourcing:** Continued access to APIs, exploring other platforms

© 2024 Institute of Data

Data Overview (1)

Stress Detection from Social Media Articles

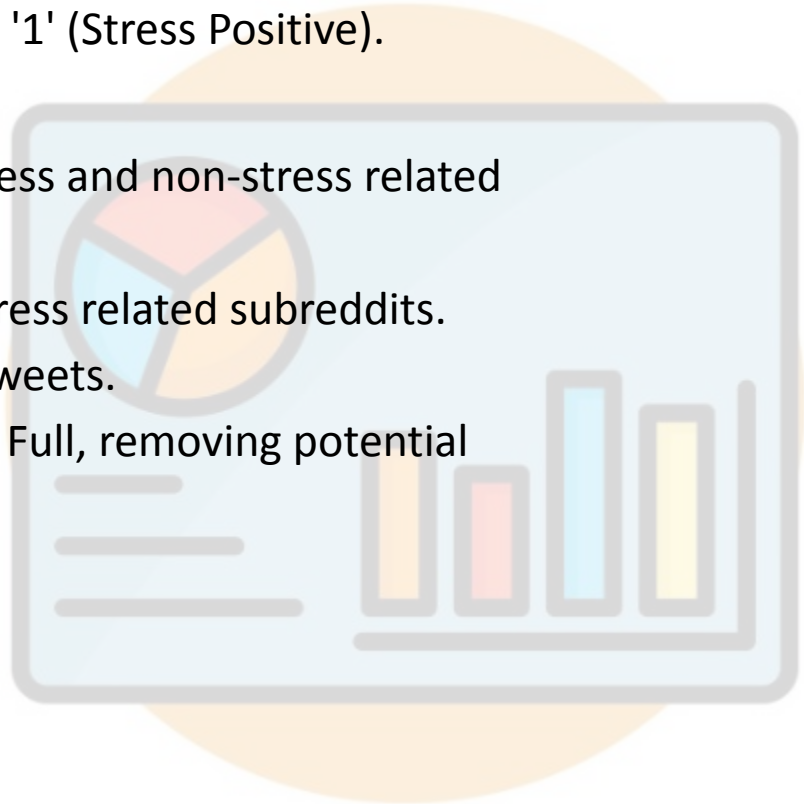
- **Source:** Kaggle Dataset
- **Objective:** Develop a more accurate and efficient NLP-based approach to automatically detect stress in social media text, enabling early intervention and improved mental health support.

© 2024 Institute of Data

Data Overview (2)

Stress Detection from Social Media Articles

- **Datasets:**
  - Constructed four datasets using text articles from Reddit and Twitter.
  - Each article is labeled with a class value of '0' (Stress Negative) or '1' (Stress Positive).
- **Dataset Descriptions:**
  - **Reddit Combi:** This dataset combines title and body text from stress and non-stress related subreddits.
  - **Reddit Title:** This dataset consists of titles from stress and non-stress related subreddits.
  - **Twitter Full:** This dataset contains stress and non-stress related tweets.
  - **Twitter Non-Advert:** This dataset is a denoised version of Twitter Full, removing potential advertisements.

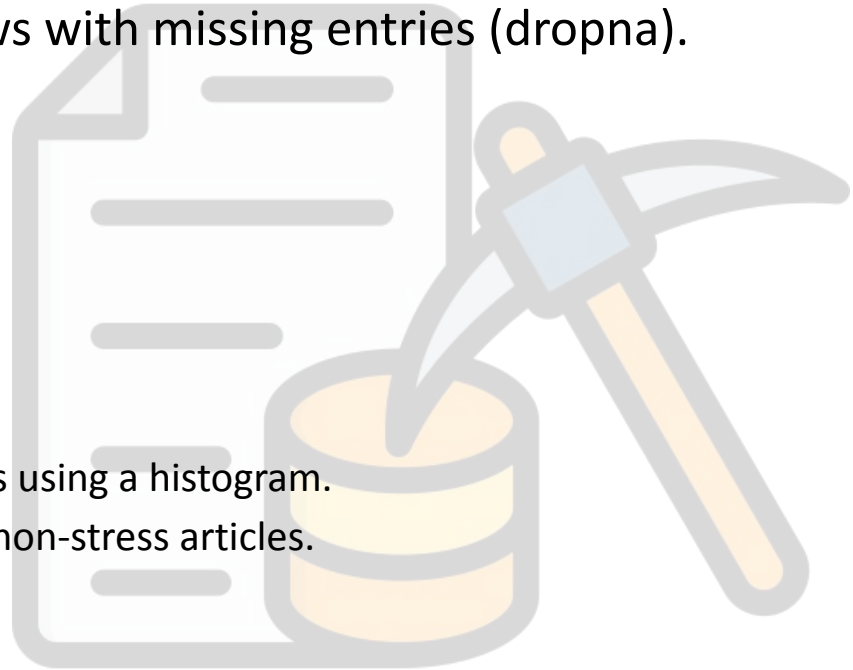


© 2024 Institute of Data

Data Exploration (1)

Understanding the Data

- Explored the data using Python libraries like pandas to understand the number of rows, column names, and data types.
- Examined the first few rows of each Data Frame to get a sense of the content and labels.
- Checked for missing values and handled them by dropping rows with missing entries (dropna).
- Preprocessed the data:
  - Dropped unnecessary columns (Reddit Combi).
  - Cleaned hashtags in Twitter data using regular expressions.
  - Concatenated all preprocessed Data Frames into a single one.
- Analyzed the data distribution:
  - Visualized the distribution of stress labels using a bar chart.
  - Examined the distribution of text length for stress and non-stress articles using a histogram.
  - Generated word clouds to visualize frequently used words in stress and non-stress articles.



© 2024 Institute of Data

Data Exploration (2)

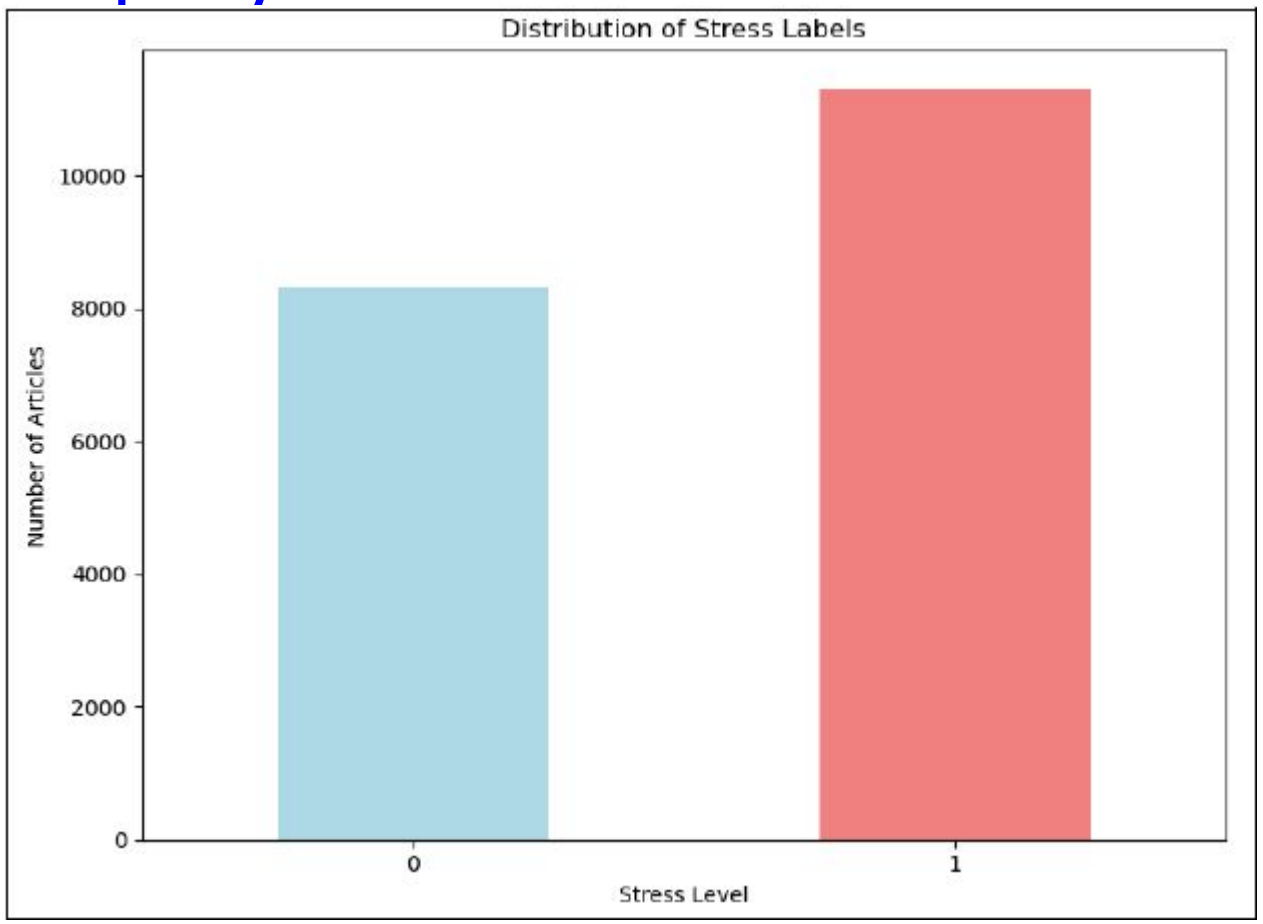
Understanding the Data

- **Number of Rows and Columns:**
  - Reddit Combi: 3123 rows, 4 columns
  - Reddit Title: 5556 rows, 2 columns
  - Twitter Full: 8900 rows, 3 columns
  - Twitter Non-Advert: 2051 rows, 2 columns
- **Data Types:**
  - **title**: object (text)
  - **label**: int64/boolean (stress label)
  - **body**: object (text) (only in Reddit Combi)
  - **hashtags**: object (text) (only in Twitter Full)
- **Missing Values:**
  - Handled missing values by dropping rows with missing entries in body column.
- **Data Preprocessing:**
  - Dropped unnecessary columns in Reddit Combi.
  - Cleaned hashtags in Twitter Full using regular expressions.
  - Concatenated all datasets into a single DataFrame.
-

© 2024 Institute of Data

Data Exploration (3)

Frequency of Stressful and Non-Stressful Posts



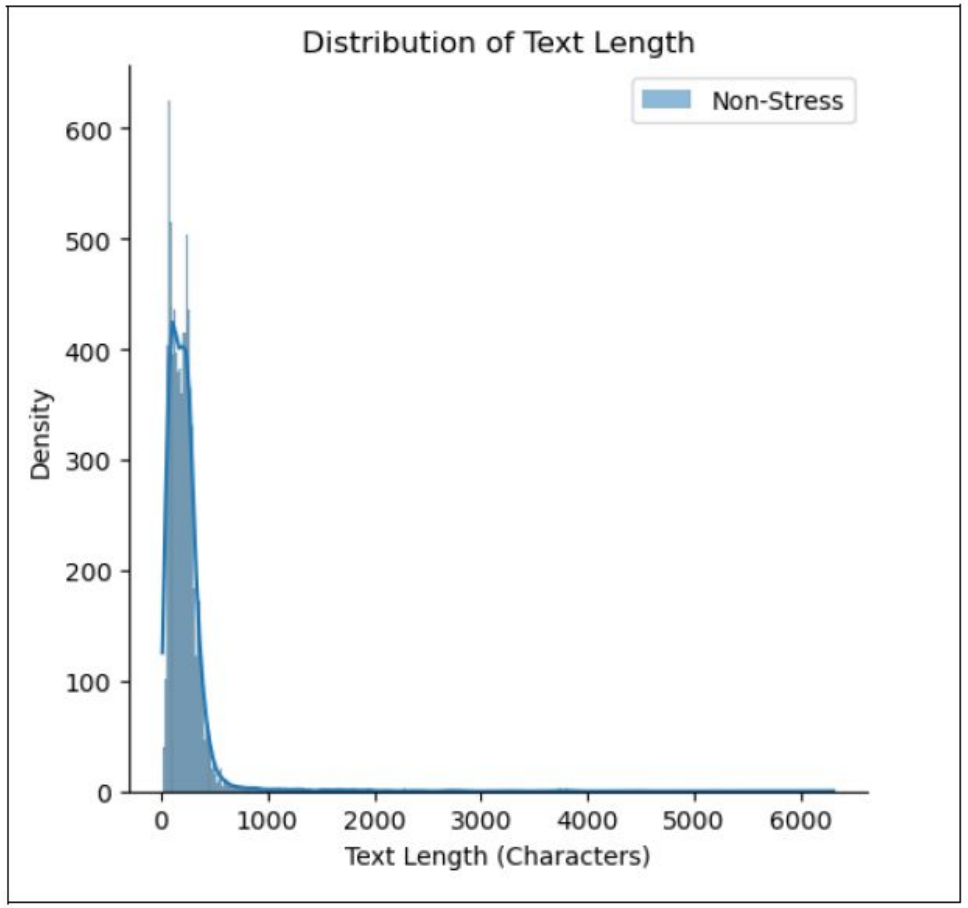
Key Notes:

- **Class imbalance:** More non-stressful articles than stressful ones.
- **Impact on modeling:** Require techniques like class weighting or oversampling.
- **Further investigation:** Explore factors contributing to imbalance (e.g., labeling difficulty, data collection bias).

© 2024 Institute of Data

Data Exploration (4)

Distribution of Text Length



Key Notes:

- **Skewed distribution:** Most posts are relatively short.
- **Overlapping distributions:** Text length alone may not be a strong predictor.
- **Non-stressful posts:** Slightly longer on average

© 2024 Institute of Data

Data Exploration (5)

Common Words Associated with Stress

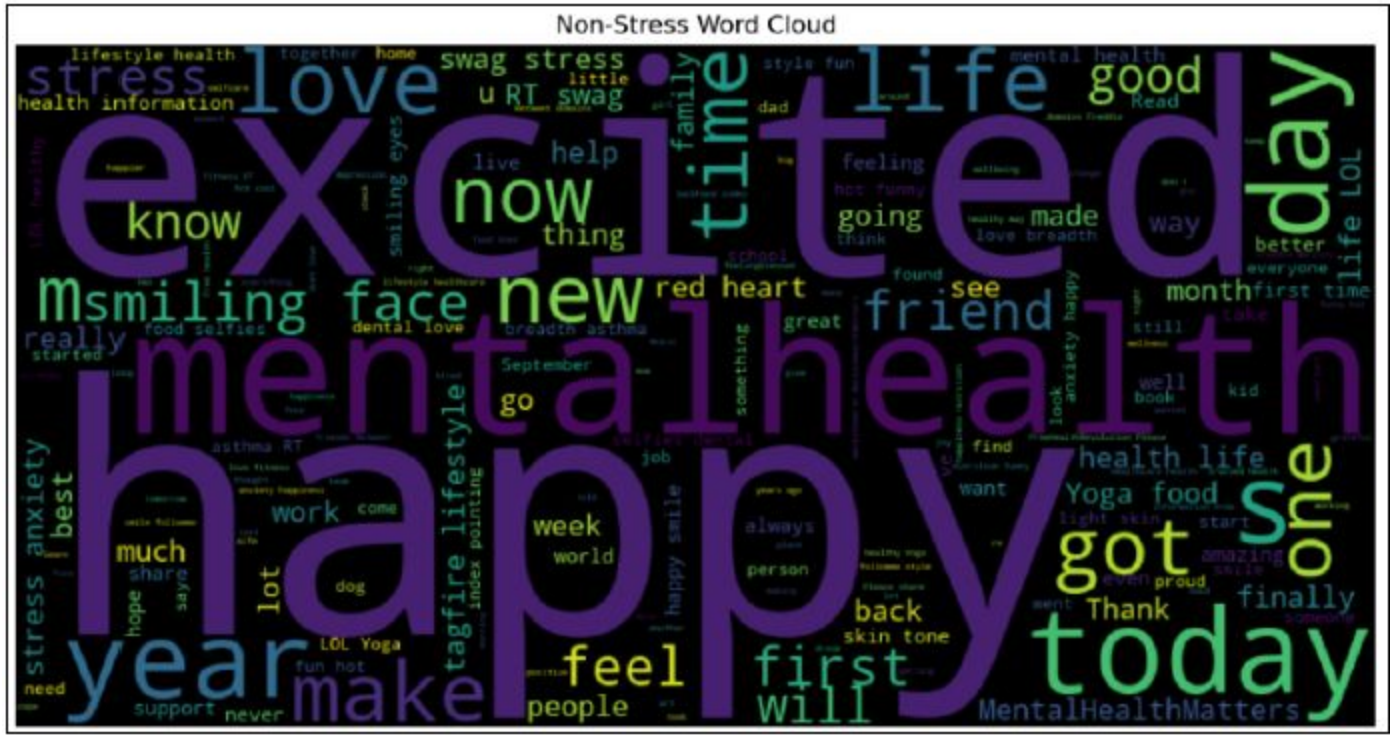
Key Notes:

- **Negative emotions:** "stress," "depressed," "sad," "anxious"
- **Life challenges:** "work," "school," "relationships," "personal problems"
- **Seeking help:** "help," "support," "therapy"

© 2024 Institute of Data

Data Exploration (6)

Common Words Associated with Non-Stress



Key Notes:

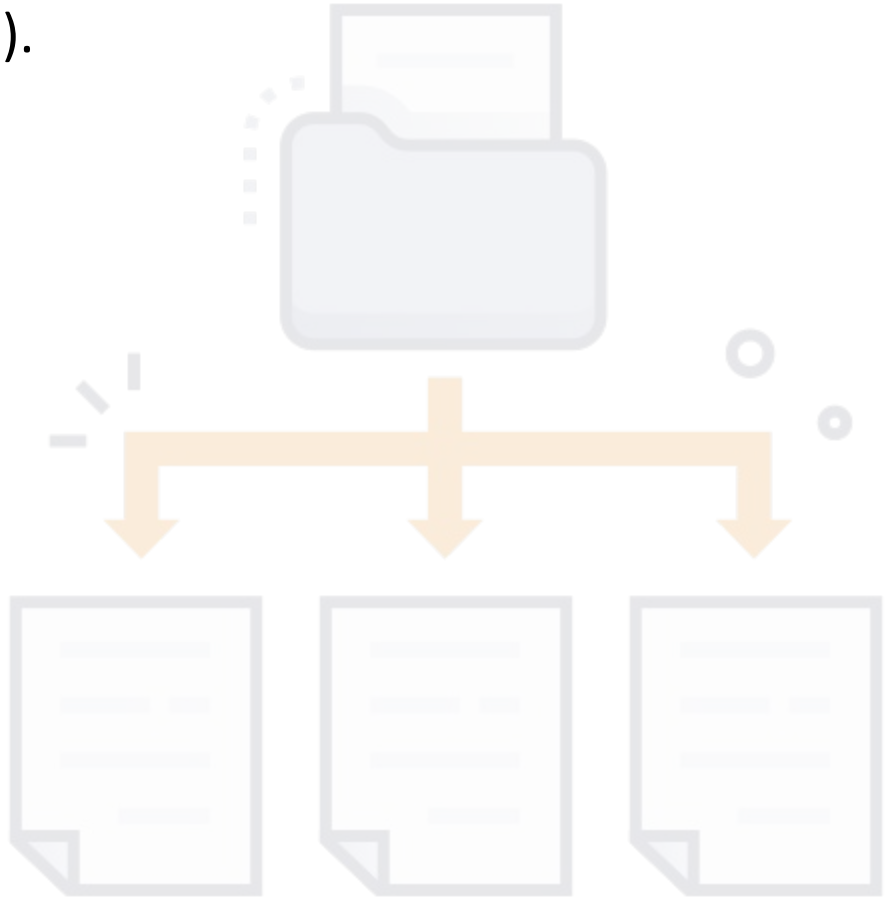
- **Positive emotions:** "happy," "love," "good," "excited"
- **Everyday life:** "daily activities," "hobbies," "social interactions"
- **Gratitude and appreciation:** "thankful," "grateful," "proud"

© 2024 Institute of Data

Data Split

Data Preparation for Modeling

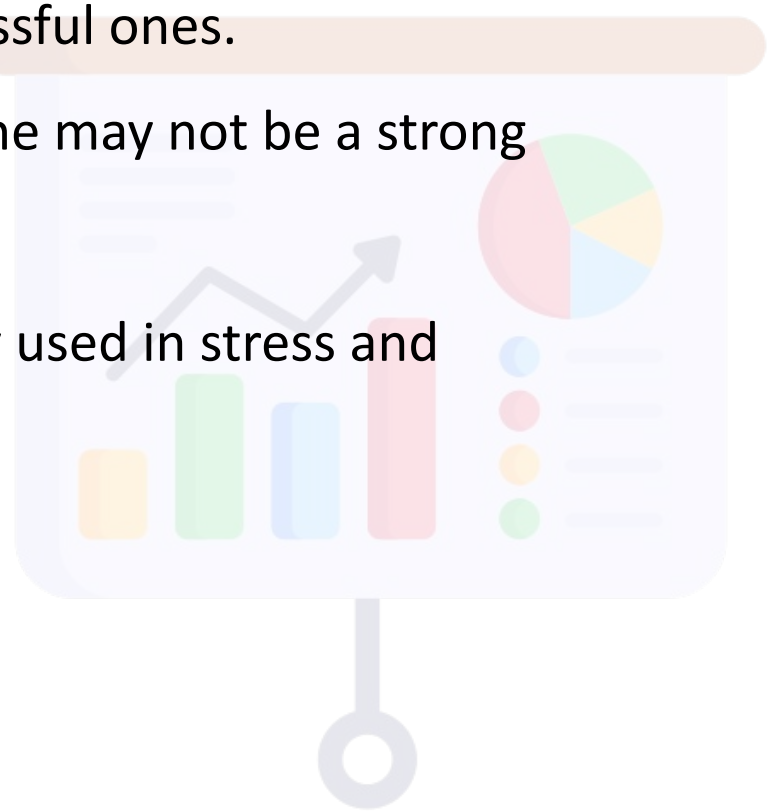
- Split data into features (title) and target (stress label).
- Split data into training (80%) and testing (20%) sets.



© 2024 Institute of Data

Data Overview

- **Key Findings:**
  - **Class imbalance:** More non-stressful articles than stressful ones.
  - **Overlapping text length distributions:** Text length alone may not be a strong predictor of stress.
  - **Distinct word patterns:** Different words are frequently used in stress and non-stress articles.



© 2024 Institute of Data

Deliver (1)

Feature Engineering

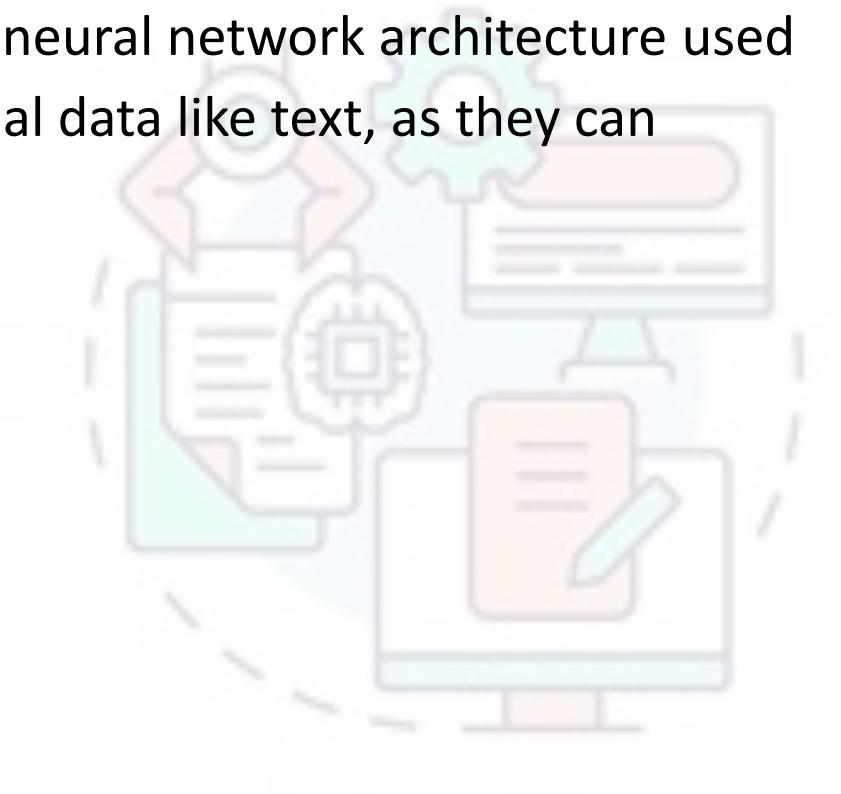
- **Key Features:**
  - Text data (primary feature)
  - Captures sentiment, emotions, and vocabulary
- **Business Significance:**
  - Text features are crucial for understanding the linguistic cues associated with stress.
  - Effective feature engineering can improve model performance and interpretability.
- **Techniques:**
  - **Text Cleaning:**
    - Removes irrelevant information (punctuation, stop words, hashtags).
  - **Text Normalization:**
    - Lowercases text for consistency.
  - **TF-IDF Vectorization:**
    - Converts text into numerical features representing word importance.

© 2024 Institute of Data

Deliver (2)

Machine Models Used

- **LinearSVC:** This is a linear support vector classifier used in the initial model. It is a good choice for text classification tasks due to its simplicity and efficiency.
- **LSTM (Long Short-Term Memory):** This is a recurrent neural network architecture used in the later model. LSTMs are well-suited for sequential data like text, as they can capture long-term dependencies between words.



© 2024 Institute of Data

Deliver (3)

Evaluation Metrics

- **Accuracy:** Measures the proportion of correct predictions made by the model.
- **Confusion Matrix:** Visualizes the number of correct and incorrect predictions for each class (stressful vs. non-stressful).
- **Classification Report:** Provides detailed information about the model's performance, including precision, recall, and F1-score for each class.
- **ROC AUC Score:** This metric is used for imbalanced datasets and measures the model's ability to distinguish between classes (AUC-ROC score closer to 1 indicates better performance).

© 2024 Institute of Data

Deliver (4)

Evaluation Metrics - Accuracy

-

© 2024 Institute of Data

Deliver (5)

Evaluation Metrics - Confusion Matrix

© 2024 Institute of Data

Deliver (6)

Evaluation Metrics - Classification Report

-

© 2024 Institute of Data

Deliver (7)

Evaluation Metrics - ROC AUC Score

-

© 2024 Institute of Data

Summary, Conclusions & Next Steps

- **Summary**
  - A brief recap of the presentation
- **Conclusions**
  - What has been achieved?
- **Next steps**
  - How can this project be developed further and implemented in real life?

© 2024 Institute of Data

Appendices

© 2024 Institute of Data

References

Data Source:

https://www.kaggle.com/datasets/mexwell/stress-detection-from-social-media-articles

Source Code:

https://github.com/jimmychong1983/SocialMediaStressDetection

© 2024 Institute of Data

Questions

© 2024 Institute of Data

Thank you
End of Presentation

© 2024 Institute of Data

Case study: Home loans marketing

Results comparison and business case overview

Applying the model for Banking can lead to potential annual **revenue twice as big** as the current model.

Results overview

	Baseline Model	Full Feature Model	Difference
% of identified applicants in top 10%	32%	61%	+29%
Potential Profit	627 x $Y	1,200 x $Y	573 x $Y

Business case overview based on the Final Model

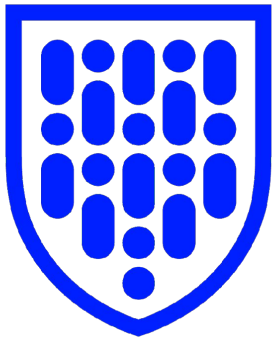
Assumptions:

- Customer Value/year is $**1000**
- Customer base = 1.4 million
- Top 10% = 140,000 customers
- 2.8% applicants over 2 years ie. 1.4% annually
- 1.4% applicants in top 10% = 1,960
- 61% identified to target = 1,200
- 32% identified to target = 627

Potential profit = 573 x 1000 ≃ $500,000/ year

≃ $1.5m over 3 years

© 2024 Institute of Data



Institute of Data

2024

© 2024 Institute of Data

Data Science and AI

Capstone Project

Harnessing NLP to Detect Stress in Social Media

Early Intervention for Mental Wellbeing

© 2024 Institute of Data

Capstone Project

- You are required to **define, design and deliver** a **data science project** towards the end of the course.
- Project milestones:
  - <date tbd> : present **3 ideas for the project**
  - <date tbd> : decide on one option
  - <date tbd> : collect data
  - <date tbd> : present initial findings
  - <date tbd> : present an update
  - <date tbd> : Dry run of final presentation
  - <date tbd> : **Present final report**

© 2024 Institute of Data

What to present

- **Business perspective**
  - Business insights uncovered
  - Business scenarios for how the project can be deployed and used
  - Approach for estimating business value
- **Technical perspective**
  - Techniques used
  - Pipeline
  - Model validation results

© 2024 Institute of Data

Project evaluation criteria

The project is evaluated on the quality, clarity and completeness of the definition, design and delivery of the project.

- **Definition (20%)**
  - Business context, stakeholders and value
  - Data description, sources, quality
- **Design (30%)**
  - Data exploration, analysis and visualisation
  - Documentation: text document, presentation and Notebook
  - The project planning, effort allocation and next steps
- **Delivery (50%)**
  - Feature Engineering
  - Creation of an effective reproducible pipeline
  - Machine Learning model algorithms and accuracy
  - Overall end-to-end solution
  - Delivery of the presentation, poise and audience engagement

© 2024 Institute of Data

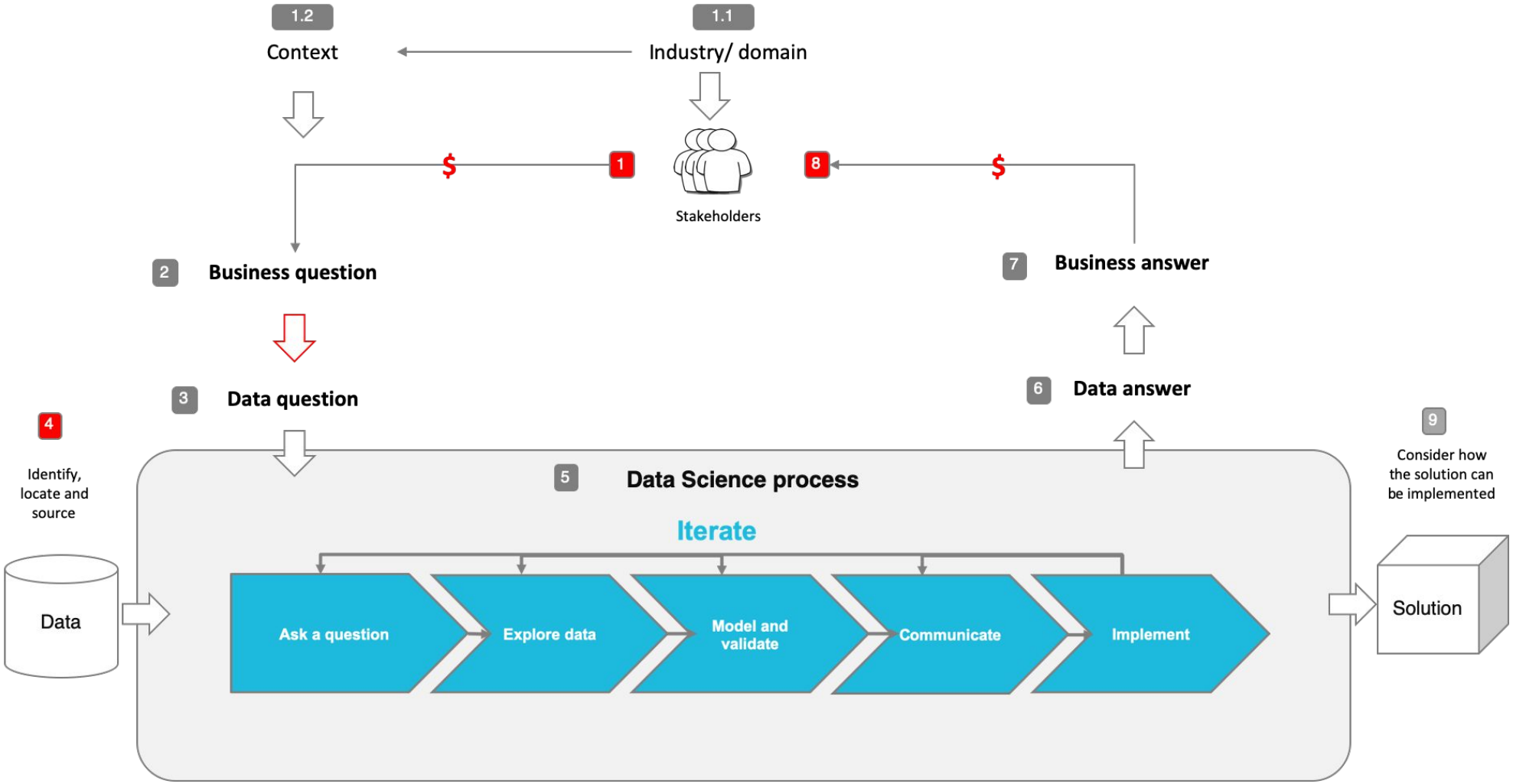
Questions?

© 2024 Institute of Data

Presentation Skeleton

© 2024 Institute of Data

Applying data science in an industry project



© 2024 Institute of Data