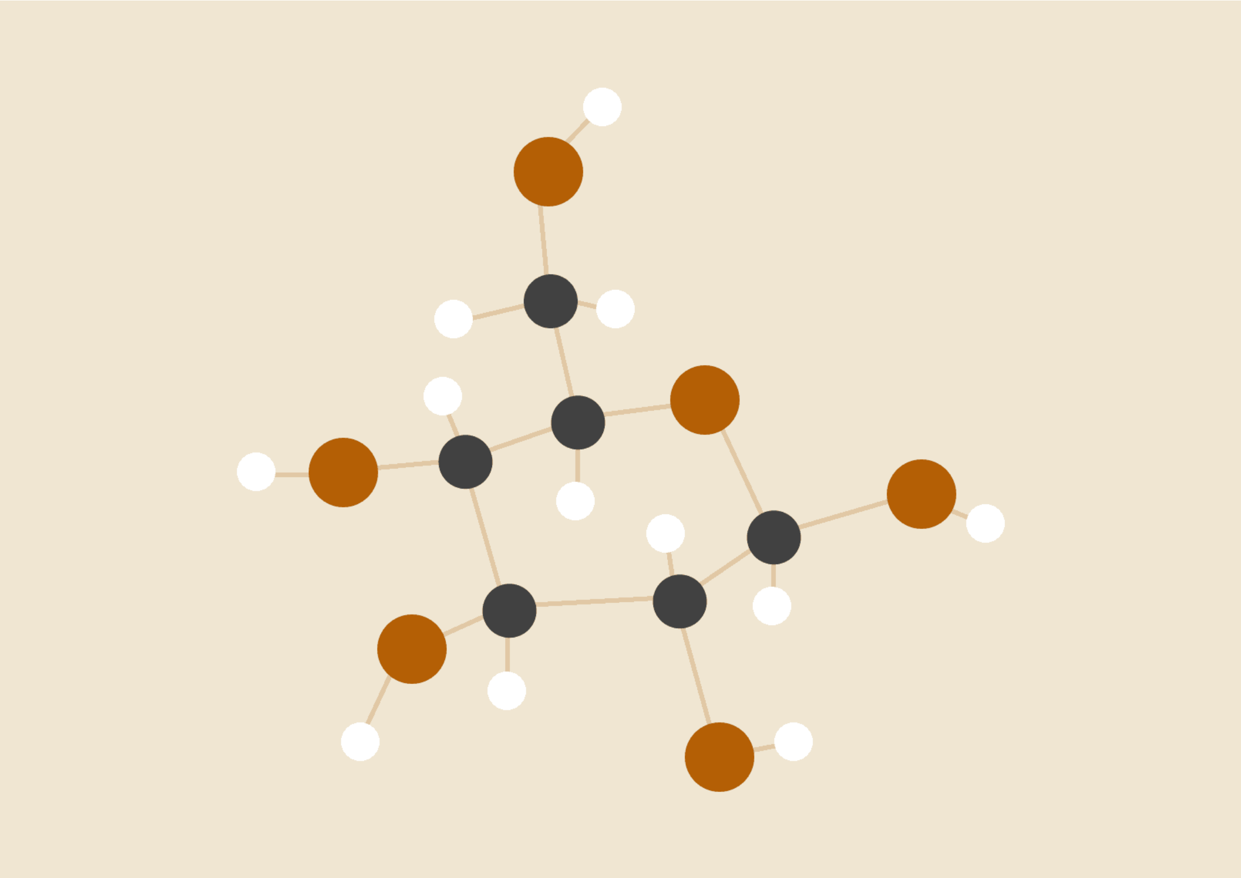
Project Proposal

AI modeling based on seed



M.MZAOUALI

April, 2025

Project Overview

We are looking for a data scientist that could produce two models to classify YouTube channels for us.

Model 1: Demographics

Build a model that would help predict demographics of YouTube channel viewers. We have raw demographics data (Gender and Age) from a couple thousands of channels and looking to model from that seed data to expand to 1MM channels. We will be able to compare the model results with channels we will use as holdout so we can assess accuracy.

Data available for each channel would include:

* Name
* Channel ID
* Channel Link
* Country
* Language
* Subscribers
* Total Views
* Number of Videos
* Last Video Title
* Channel Tags
* Channel Description
* Category
* Top 25 Video Tags
* Profile Picture
* Made for kids flag

Model 2: Contextual

We also want a model that is able to classify channels based on a detailed list of contextual segments (e.g Sports, Finance) based on the channel information above. We will provide the list of context to match to.

Project Plan

As the project contains two sets of approach, prediction and segmentation, it is best to my knowledge, to incorporate the best of the two worlds of supervised and unsupervised modeling.

The project can be executed, for a 30h work week, within 12 to 15 weeks.

Here's a comprehensive 12-week project plan incorporating:

both supervised and unsupervised learning approaches.

with detailed tasks per week.

Including key deliverables.

12-WEEK PROJECT PLAN

1. PHASE 1: FOUNDATION (Weeks 1-3)
   a. Week 1: Data Collection and Initial Analysis
      i. - Review and clean the provided dataset
      ii. - Perform exploratory data analysis (EDA)
      iii. - Handle missing values and outliers
      iv. - Set up development environment and version control

   Deliverable: Clean dataset and EDA report

   b. Week 2: Feature Engineering (Text & Metadata)
      i. - Text preprocessing (descriptions, tags, titles)
      ii. - Convert channel tags and video tags into features
      iii. - Create numerical features from subscriber/view ratios
      iv. - Develop language-specific features

   Deliverable: Initial feature engineering pipeline

   c. Week 3: Feature Engineering (Advanced)
      i. - Extract features from profile pictures using computer vision
      ii. - Implement text embedding techniques (BERT/Word2Vec)
      iii. - Create cross-feature interactions

   Deliverable: Complete feature engineering pipeline

2. PHASE 2: DEMOGRAPHICS MODEL (Weeks 4-6)
   a. Week 4: Demographics Model - Initial Development
      i. - Split data into training and validation sets
      ii. - Build initial classification models for gender prediction
      iii. - Implement basic age group classification

   Deliverable: Initial demographics models

b. Week 5: Demographics Model - Enhancement
i. - Optimize model hyperparameters
ii. - Implement ensemble methods
iii. - Enhance age group classification accuracy

Deliverable: Optimized demographics models

c. Week 6: Demographics Model - Finalization
i. - Validate against holdout set
ii. - Fine-tune models based on validation results
iii. - Document performance metrics

Deliverable: Final demographics prediction system

3. PHASE 3: CONTEXTUAL ANALYSIS (Weeks 7-9)
a. Week 7: Unsupervised Learning - Base
i. - Implement Topic Modeling using LDA
ii. - Generate BERT embeddings for channels
iii. - Apply dimensional reduction (UMAP/t-SNE)
iv. - Create initial clustering models (K-means, DBSCAN)

Deliverable: Base unsupervised learning models

b. Week 8: Unsupervised Learning - Advanced
i. - Implement hierarchical clustering
ii. - Create interactive cluster visualizations
iii. - Develop topic distribution analysis
iv. - Generate cluster insights report

Deliverable: Advanced clustering analysis

c. Week 9: Supervised Contextual Classification
i. - Build supervised classifier using provided categories
ii. - Integrate cluster insights
iii. - Implement hybrid classification approach
iv. - Create confidence scoring system

Deliverable: Hybrid contextual classification system

4. PHASE 4: OPTIMIZATION & SCALING (Weeks 10-12)
a. Week 10: System Integration
i. - Combine demographics and contextual models
ii. - Implement batch processing pipeline
iii. - Optimize for 1MM channel processing
iv. - Set up efficient prediction pipeline

Deliverable: Integrated classification system

b. Week 11: Testing and Validation
    i. - Comprehensive testing with holdout data
    ii. - Performance optimization
    iii. - Error analysis and handling
    iv. - Fine-tune based on client feedback

    Deliverable: Validated and optimized system

c. Week 12: Documentation and Deployment
    i. - Create detailed technical documentation
    ii. - Prepare deployment guidelines
    iii. - Knowledge transfer sessions
    iv. - Final presentation preparation

    Deliverable: Final system with complete documentation

CRITICAL PROJECT REQUIREMENTS

To ensure:

- Clear project boundaries
- Efficient communication
- Quality deliverables
- Measurable outcomes
- Successful project completion

Some additional requirements are to be met:

A. Data Access Requirements:

1. Complete Dataset
   - Raw demographics data for initial channels
   - All channel metadata (subscribers, views, etc.)
   - Text content (descriptions, tags, titles)
   - Visual content (profile pictures)
   - Historical performance metrics if available

2. Classification Framework
   - Comprehensive list of contextual segments/categories
   - Clear category definitions and hierarchies
   - Examples of channels for each category
   - Any existing classification guidelines
   - Category overlap rules (if applicable)

3. Validation Data
   - Access to holdout dataset
   - Validation criteria
   - Acceptable accuracy thresholds
   - Known edge cases
   - Example of correctly classified channels

B. Project Communication Requirements:

1. Regular Check-ins
   - Weekly progress review meetings
   - Milestone validation sessions
   - Technical consultation as needed
   - Performance review discussions
   - Stakeholder feedback sessions

2. Documentation Requirements
   - Data dictionary
   - Classification guidelines
   - Business rules documentation
   - Success criteria definition
   - Reporting format preference

C. Timeline Dependencies:

- Initial data delivery date
- Category framework sign-off
- Regular feedback schedule
- Validation data availability
- Final delivery expectations

Budget Estimation

- 12 weeks of full-time work (30 hours/Week for a $10.00 USD per hour)

- Suggested cost range: $1000.00 - $1200.00 USD

Conclusion

This 12-week plan provides more time for:

- Thorough model development
- Extensive testing and validation
- Better integration of unsupervised insights
- Comprehensive documentation
- Client feedback incorporation

Would you like me to elaborate on any specific phase or component of this plan?

Let’s discuss this proposal in detail through a video call.