CareerVillage Questions Matching Recommendation System

Presented By: Minh Duong

Context & Problem Statement

- CareerVillage.org, a non-profit for career guidance, aims to boost impact through a data-driven recommendation system.
- The system connects students' questions with suitable volunteers, enhancing connections between aspiring individuals and experienced professionals.
- Leveraging data science, the project optimizes engagement, ensuring students receive valuable advice, empowering underserved youth with career role models.
- The CareerVillage Question Recommendation Model efficiently matches students' questions with professionals, outlined in this report covering methodology, data exploration, model development, and key findings.

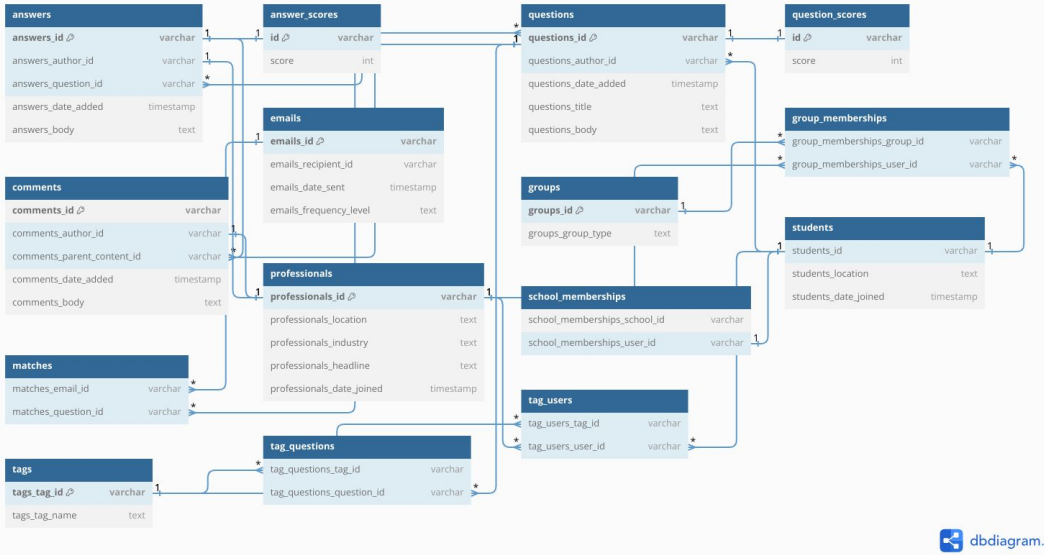
Methodology

Content-based Approach

- Privacy is a top priority; no personal or usage data is stored or used for training in the recommendation system.
- The project opts for a content-based approach to address privacy concerns.
- Two recommendation models were created, building professional profiles from past answered questions and tags.
- These models were evaluated based on recommendation success rates, comparing favorably to existing approaches.

Data

Source: Kaggle-provided data from CareerVillage.



Data

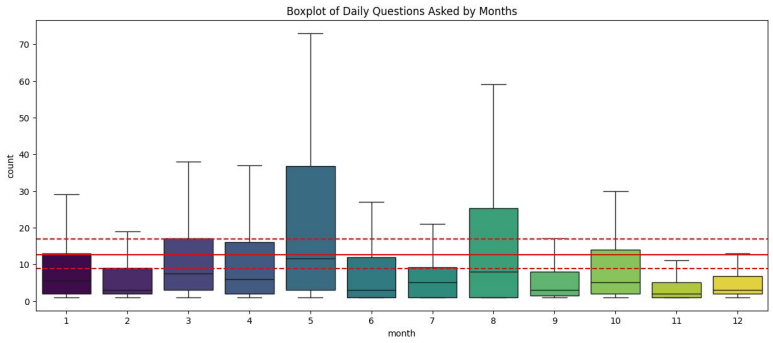
Key Tables:

- Questions: 23,931 records
- Professionals: 28,152 records
- Answers: 23,110 records
- Emails:
- Matches:

Exploratory Data Analysis

Key findings:

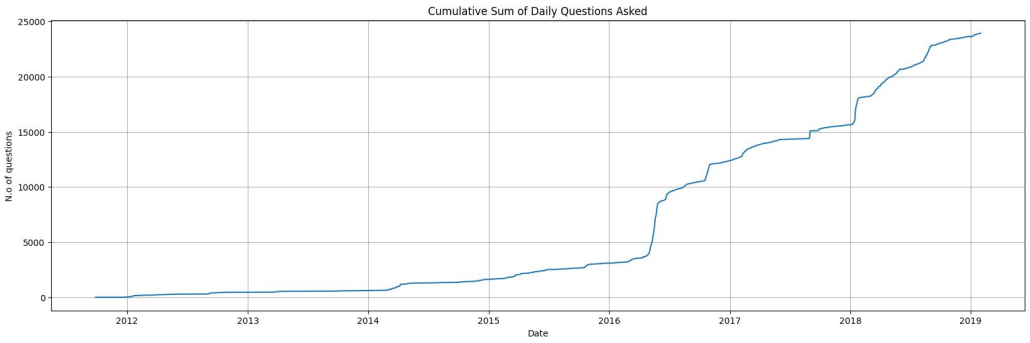
- Peak in daily questions during graduation months.



Exploratory Data Analysis

Key findings:

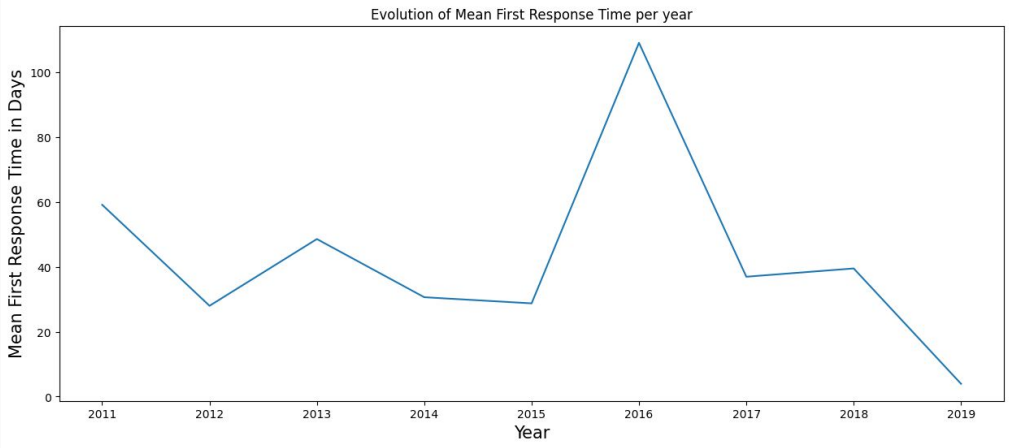
- Significant increase in questions since 2016



Exploratory Data Analysis

Key findings:

- Average response time for questions is 65 days.



Data Cleaning & Preprocessing

- Most time-consuming step.
- Multiple tables required merging and examination.
- Goal: Understand table relations and create data for modeling.

Data Cleaning & Preprocessing

Train/Test Split:

- Unconventional split based on midpoint of answers table by date.
- Train set: Professional IDs, combined content of past questions, list of tags for each professional.

Data Cleaning & Preprocessing

Professionals and Questions Profiles:

- Merged questions and answers for profiles in the train set.
- Created a question_tag_list table with tags associated with each question.
- Merged professionals with tag_users to create tags feature.
- Combined data frames for professional profiles: professional IDs, tags, content of past answered questions

Data Cleaning & Preprocessing

Test Set Profiles and Text Cleaning:

- Applied the same merging technique for profiles in the test set.
- Text cleaning function: Removed special characters, stopwords, digits; lemmatized and converted text to lowercase.

Data Cleaning & Preprocessing

Handling Missing Values:

- Two approaches considered for missing tags.
- Chose to impute missing tag values with "no_tag" for simplicity.
- Second approach (auto-tag-generated function) considered for future improvement.

Modeling - Question-Content Based

TF-IDF Vectorization

- Transformed textual content of train set questions into numerical vectors using TF-IDF.
- Assigned weights to words based on importance, creating a numerical representation.

Modeling - Question-Content Based

Model Training:

- Trained the model using TF-IDF vectors and professional response data.
- System learned relationships between question content features and professional preferences.

Modeling - Question-Content Based

Recommendation Generation:

- Assessed content similarity using cosine similarity.
- Identified questions with similar attributes based on historical preferences.
- Generated recommendations of 10 professionals for each question.

Modeling - Tags-based

- Built a similar model using the same techniques as the content-based model.
- Employed tags of professionals and questions to calculate cosine similarity.

Model Evaluation - Evaluation Metrics

- Establishing a nuanced matching score for professionals and questions.
- Criteria for success include Answer Rating, Response Time, and Response Rate.
- Emphasizes the inadequacy of the current answer rating system during EDA.
- Advocates for a more sophisticated rating system for robust evaluation.

Model Evaluation - Response Rate and Time Metrics

- Response Rate: Ratio of responses to total assigned questions, gauges engagement.
- Response Time: Duration to respond plays a crucial role in efficiency assessment.

Model Evaluation - Evaluation Process

- Initial focus on utilizing Response Rate as a predominant metric.
- Determining original response rates of questions in the test set.
- Merging emails and matches to identify assigned questions.
- Calculating the average response rate of the original matching system.

Model Evaluation - Model-Specific Response Rates

- Determining response rates with question-content-based recommendation.
- Model recommends questions to 10 relevant profession
- Calculating the average response rate for the content-based recommendation mod

Model Evaluation - Model-Specific Response Rates

- Determining response rates with tags-based recommendation.
- Using the same method as the content-based model.
- Calculating the average response rate for the tags-based recommendation model.

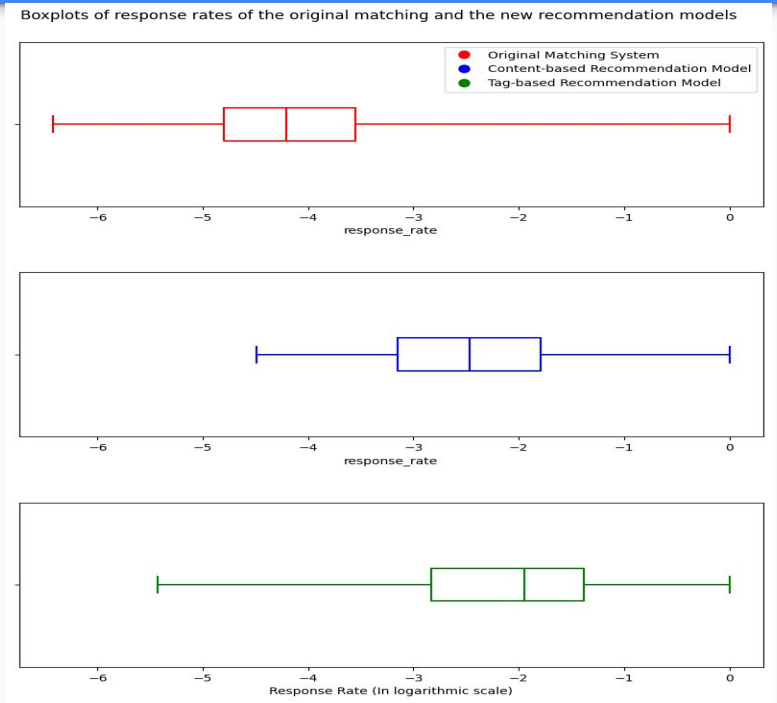
Model Evaluation - Comparison of Average Response Rates

- Involves creating boxplots for original system and two recommendation models.
- Comparison includes question content-based recommendation and tags-based recommendation.
- Rates transformed to logarithmic scale for clarity.

Model Evaluation - Comparison of Average Response Rates

Average response rate

- **Original matching system: 0.38%**
- **Question content-based recommendation model: 0.42%**
- **Tags-based recommendation model: 0.60%**



Future Steps

- Integrate response time into the success metric.
- Consider professional activeness as a feature.
- Include professionals' previous responses in training.
- Implement dynamic tagging for evolving expertise.
- Establish a user feedback loop and improve user rating mechanism.
- Explore advanced collaborative filtering techniques.

Project Impact

Significantly enhances the CareerVillage platform, empowering students with career guidance and fostering meaningful connections.

Recommendations and insights provide a solid foundation for future enhancements, ensuring a more personalized and impactful user experience on CareerVillage.org.

Thanks!

college
career
job
high school
major
one
good
wondering
university
going
important
feel
thing
time
something
want become
year college
think
kind
make
interested
best way
start
area
experience
field
work
business
help
people
need
much
psychology
dont know
internship
education
know
course
love
degree
class
technology
working
take
life
engineering
best
study
want
want know
student
future
program
better
well
sport
graduate
hard
apply
biology
etc
medical school
computer science
new
might
money
possible
idea
difficult
finance
great
school
medicine
skill
looking
interest
someone
learn
lot
see
able
come
id
use
scholarship
pursue
focus
doctor
research
go college
math
person
teacher
im interested
professional
choose
first
want go
option
bachelor degree
company
minor
dont want
im sure
even
different
go
nursing
benefit
thank
type
career path
general
may
right
look
many
science
getting
actually
advice
taking
seem
give
industry
family
music
studying
plan