3D Heals Final Report:
Webinar Attendance

Table of Contents

Background

3DHeals "aims to bridge the knowledge and experience gap between 3D printing, an emerging technology, and the established healthcare and life science innovation ecosystem." 3DHeals wants to know who is joining their free webinars in order to tailor the webinars to encourage more participants, as well as more engagement during the events. https://3dheals.com/about/

Problem Identification

The Question: What can be developed that informs 3DHeals who is joining their webinars, how they can increase the number of participants, and how they can increase engagement during the webinars?

The Audience: The results will be presented to the CEO and founder of 3DHeals Jenny Chen.

The Data: The data is from Zoom of four different webinars (Indian Ecosystem, Post-processing, Biomaterials, and 3D Metal) put on by 3DHeals in 2019.

The Time Frame: February 2023-April 2023

The Modeling Response: the percentage (between 0 and 1) of time a participant will spend in a webinar

The Model: Supervised regression using Gradient Boosting

The Deliverables: Colab notebooks of data wrangling through modeling, as well as this final report and presentation

Data Wrangling and Preprocessing

The data began as four separate csv files that Zoom provided to 3DHeals at the end of each webinar.

The original columns were as follows:

- Attended: Did the participant attend the webinar? (yes=1 or no=0)
- User Name (Original Name)
- First Name
- Last Name
- Email
- City
- Country/Region
- Zip/Postal Code
- State/Province
- Industry: what industry do you work for? (drop down menu)
- Organization: what is the name of the organization you work for? (free response)
- Job Title: what is your job title? (free response)
- Questions & Comments: do you have any questions or comments?
- Registration Time
- Approval Status
- Join Time: the time the participant joined the webinar in Pacific Daylight Time
- Leave Time: the time the participant left the webinar in Pacific Daylight Time
- Time in Session (minutes): the time between join and leave time in minutes
- Is Guest
- I agree that I may be video recorded and acknowledge that I am aware that the recording will be available to the public.
- I agree that 3DHEALS can share my information with this event's sponsor and partners (who made this event Free to me.) I can unsubscribe at any time.
- Country/Region Name

- Source Name: where did you hear about this webinar?

Columns that were created and added to the dataset:

- webinar_name: which webinar was registered for?
- Job Title Category: I condensed the job titles down into 11 groups
- Total_time: the total time a person spent in the webinar
- Number_of_Logins: the number of times a participant logged into the webinar
- Registered_Webinars: the number of webinars a participant registered for
- Attended_Webinars: the total number of webinars a participant attended
- Attended_Percent: the percentage of webinars a participant attended out of the number registered for
- Time_in_session_percent: the percent of time (between 0 and 1) that a participant spent in a webinar
- Country_timezone: the first timezone of the country
- timestamp: the join time in the participant's timezone
- night_time: if a participant logged in late at night in their timezone (yes=1 or no=0)

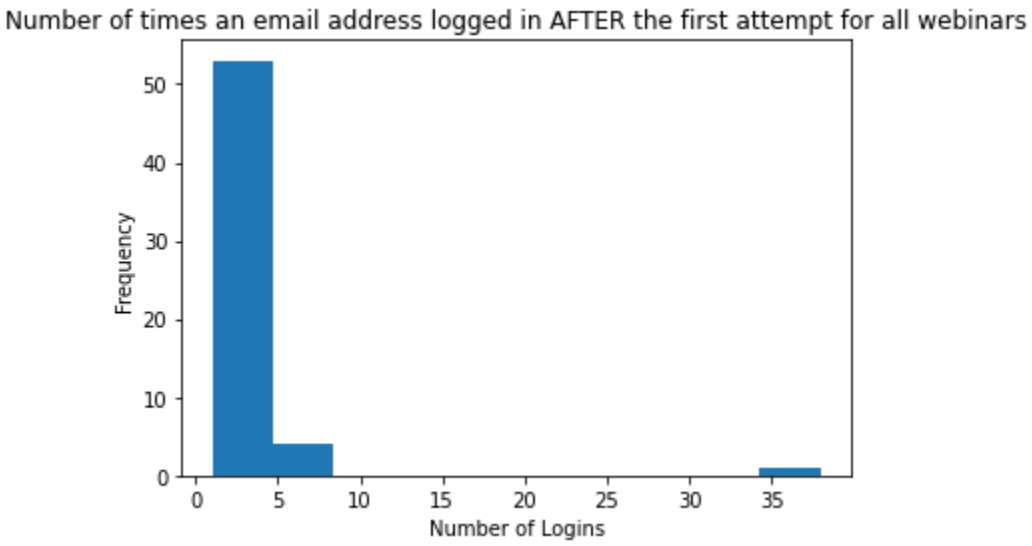
I began the data wrangling process by checking for missing values and dropping unnecessary columns. I dropped the names, the acknowledgements, the zip code because it was mostly empty, and "Is Guest" and "Approval Status" because they didn't have any meaning.

Then I checked the data type for each column. I converted the time columns to datetime type. I cleaned up the Time in Session so that those who didn't attend the webinar would have a time of zero instead of "--".

The biggest task of data wrangling was to condense the job titles down into manageable categories. Jenny wanted to be able to offer a drop down choice for job titles, instead of having the participant typing in their job title by hand. After lower casing and removing punctuation, I had 259 unique job titles. I used fuzzywuzzy with the token set ratio because it is more flexible in its matches. I chose cutoff scores of 60-80 based on what matches were the most accurate. With Jenny's guidance, I was able to condense the titles down to 11 groups, which covered 93% of the data. There were many people who didn't put their title. This won't be an issue in the future because they will be forced to make a choice from the drop down menu. See groups below.

None	382
Academic	93
C-Suite	68
Researcher	62
Manager, Director	57
Engineer	49
Clinician	22
Marketer, Salesperson	16
Designer	12
Business Developer	11
Consultant	9
Operator	2

The next task was cleaning up the duplicate values. Every time a person logged out of a webinar and logged back in, another entry was made. The subsequent entries were mostly full of missing data except the email, join, leave, and session time. I aggregated all the entries per email per webinar, keeping the demographics from the original registration row, summing the times in session, and keeping the first join time and last leave time. I also kept track of the number of times each person logged in and saved it to a new column.



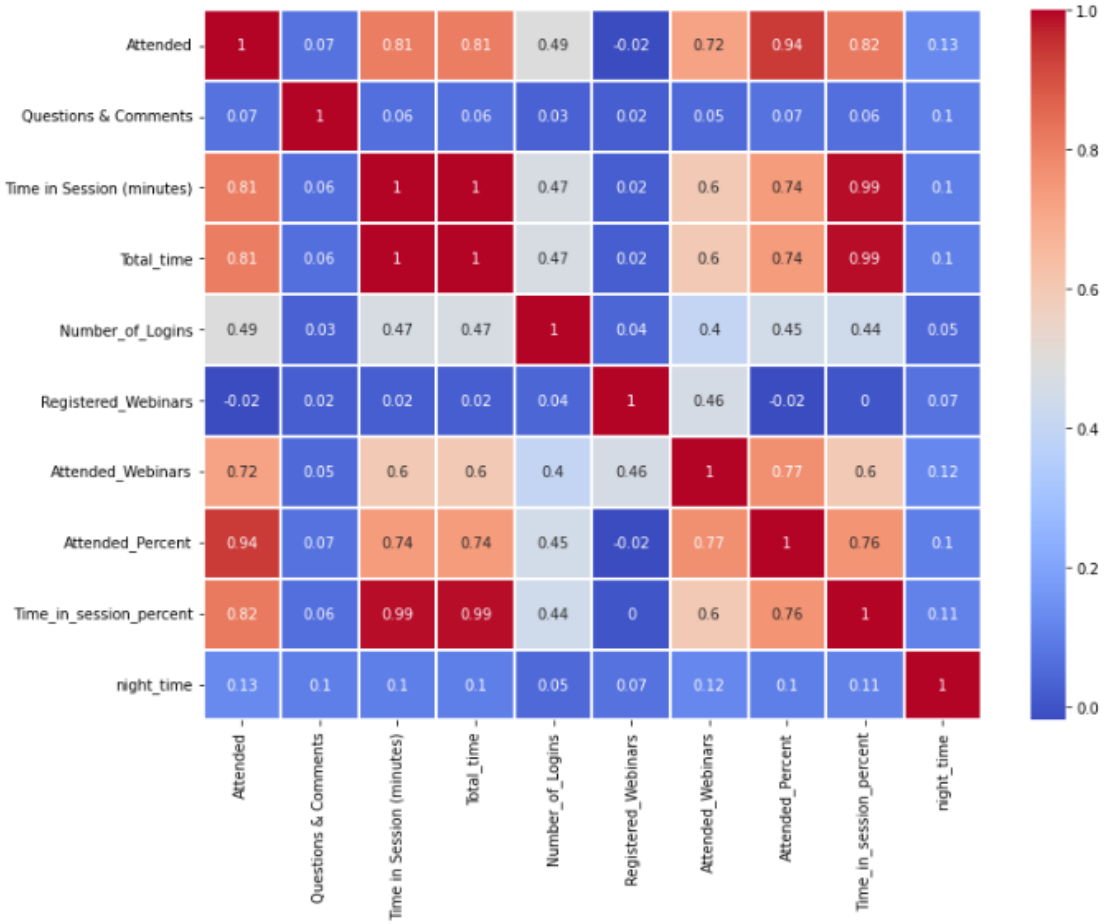
This aggregation decreased the number of missing values significantly. To finish the missing values clean up, I dropped any rows that had more than 50% missing values. I filled in 0 for anyone who did not ask a question during registration and a 1 for those that did. I cleaned up the source and imputed the mode for the missing source values. I also mapped the Attended column to No=0 and Yes=1. After I filled in missing countries (which was only 2) and missing organization (which was only 6) with "Other,"

the only missing values were the join and leave times from those who didn't join the webinar (only registered).

I then turned to feature engineering. I created a column for the number of webinars a participant registered for, a column for the total number of webinars a participant attended, and a column for the percentage of webinars a participant attended out of the number registered for. My partner Diana created a column for the percent of time a participant spent in each webinar. She also created a timezone column based on the participant's country. She was then able to determine if a person joined at night time, which could mean extra interest.

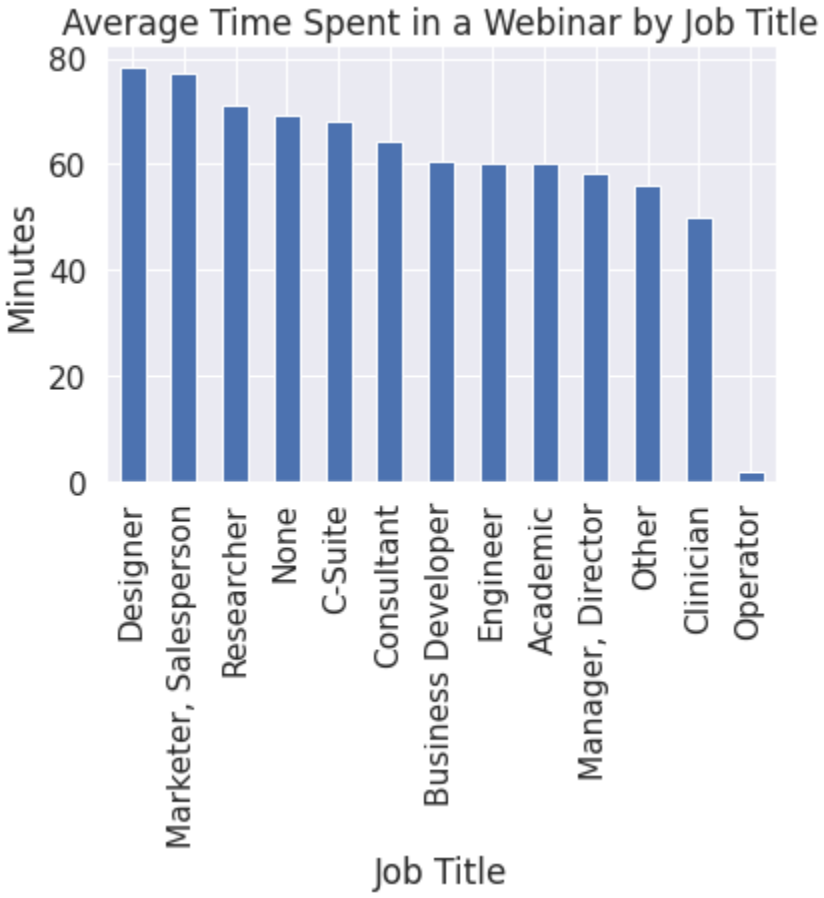
Lastly, we saved the aggregated data with new features to a new csv file to be used for Exploratory Data Analysis (EDA).

Then I spent time understanding each column by looking at the value counts of each category and creating histograms, box plots, and bar graphs where appropriate while Diana looked more closely at correlations. See below.



During Exploratory Data Analysis (EDA), I found that of those who registered the top job title categories were academic (13%), c-suite (10%), researcher (9%), manager/director (8%), and engineer (7%). Sixty percent of those who registered came from Mailchimp and 24% from LinkedIn. The main industries were medical/pharma/biotech (34%), education (13%), and hospital/clinic/doctor office (8%). The top countries were the US (31%) and India (17%).

Of those who attended (45% of those who registered), 82% attended one webinar while 12% attended two webinars. On average, designers (78 minutes), marketers/salespeople (77 minutes), and researchers (71 minutes) spent the most time in the webinars.



For preprocessing, Diana removed the columns that were highly correlated. I created indicator features for the categorical variables by using pandas get_dummies, and I removed the original columns. Y was the percentage of time a person spent in a webinar, and X was all the remaining features. Then I split the data into an 80/20 train/test. Lastly, I used StandardScaler to standardize the magnitudes of the numeric features (Number_of_Logins, Registered_Webinars, Attended_Webinars).

Model Description

I trained and cross validated 11 models: Linear Regression, Lasso, Elastic Net, KNN, Decision Tree, Gradient Boosting, Stochastic Gradient Descent, SVR, Bayesian Ridge, Kernel Ridge, and XG Boost. I only used the default parameters at first. I scored them by the negative mean squared error and converted this to root mean squared error (RMSE). We took the top performers and worked on tuning them. See below.

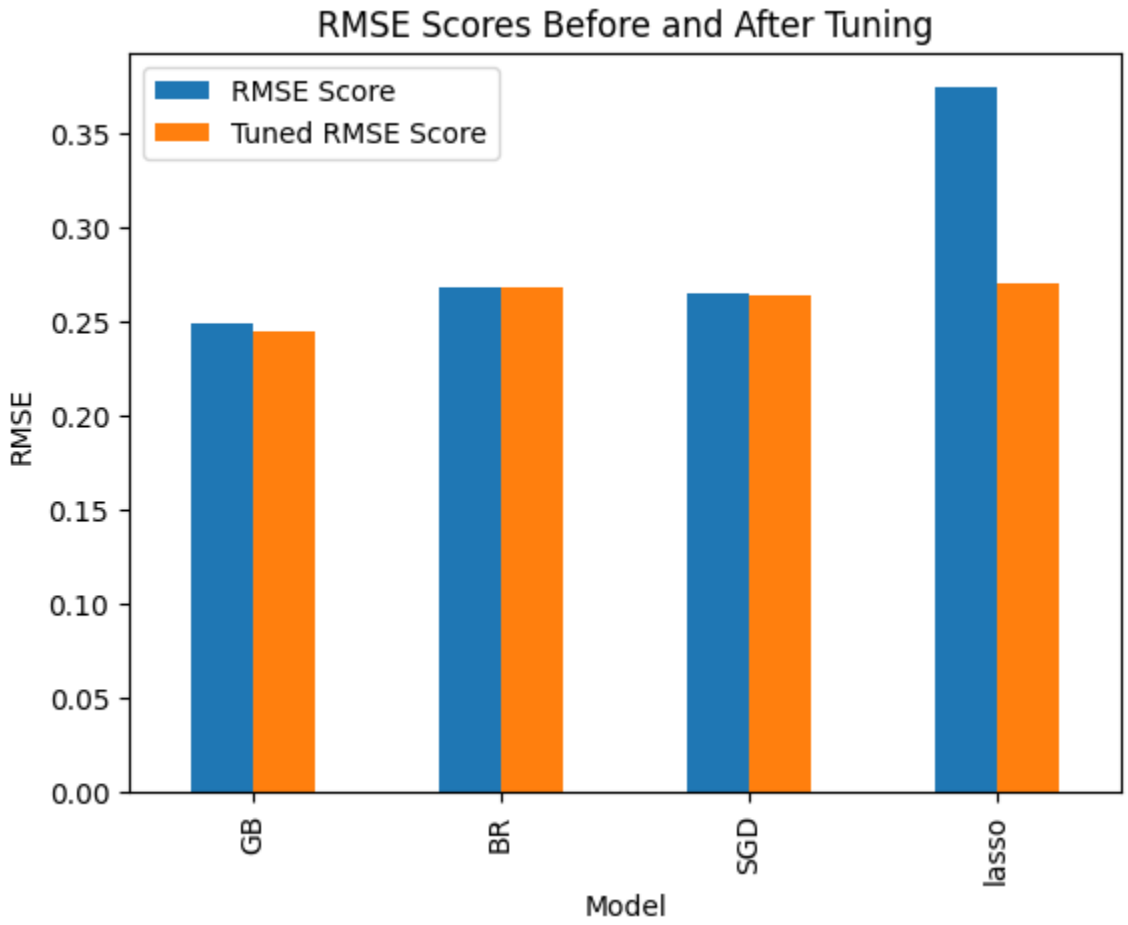
model_name	RMSE
ScaledGB	0.218149
ScaledXGBOOST	0.22738
ScaledTREE	0.287081
ScaledBR	0.293509
ScaledSVR	0.299481
ScaledSGD	0.312167
ScaledKNN	0.364155
ScaledLASSO	0.392866
ScaledEN	0.392866
ScaledKR	0.463283
ScaledLR	44452819299278.515625

Model Performance

We used the RMSE score because it measures the distance between the predicted percentage of time and the actual percentage of time a participant spent in a webinar. “In other words, it tells you how concentrated the data is around the line of best fit. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable.” (https://www.kaggle.com/general/215997)

My job was to improve the RMSE score of the Gradient Boosting, Bayesian Ridge, Stochastic Gradient Descent, and Lasso models.

Model	RMSE Score	Tuned RMSE Score	Difference in Scores
GB	0.248724	0.244326	-0.0044
BR	0.268594	0.268594	0.0000
SGD	0.264991	0.263533	-0.0015
lasso	0.373767	0.270498	-0.1033



The best parameters were as follows:

```
Gradient Boosting: {'n_estimators': 50, 'min_samples_split': 10,
'max_depth': 10, 'learning_rate': 0.1}

Bayesian Ridge: {'n_iter': 500, 'lambda_init': 0.1, 'alpha_init': 1.5}
```

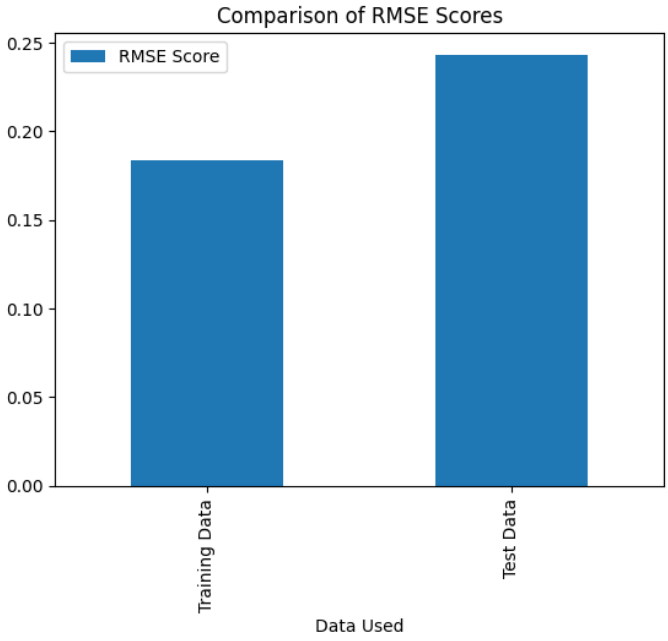
```
Stochastic Gradient Descent: {'max_iter': 5, 'epsilon': 1e-06, 'alpha':
0.01}

Lasso: {'max_iter': 500, 'alpha': 0.01}
```

Each model showed improvement over the dummy regressor (predicting the median for each participant.

Model	Dummy Difference
GB	0.208451
BR	0.187268
SGD	0.194733
lasso	0.185364

To ensure Gradient Boosting was not overfitted, I compared the training RMSE to the test RMSE and found a slight difference.

Data Used	RMSE Score
Training Data	0.183812
Test Data	0.243263



Next Steps

3DHeals has 40+ more datasets of different webinars that could be analyzed to improve the accuracy of the model and to glean more information. More demographic data would be helpful as well because the current data does not include many features. It

could be informative to know more about each participant (gender, age, income, etc). Including more information for each webinar would also be interesting, such as who was the speaker and what topics were included. Future questions could involve exploring who asked a question in the chat and performing clustering on the attendees.

To use the model to its fullest, the model will need to be deployed. A data scientist could then clean and input data into the model to see what percentage of the webinar will be attended by each person.