LEARN PLATFORM BY INSTRUCTURE

FINAL REPORT

EdTech Synergy Innovators (Team 53)

Kay Quiballo, Rohan Tandon, Lauren Vaught, Ge Li, Brandon Eubank

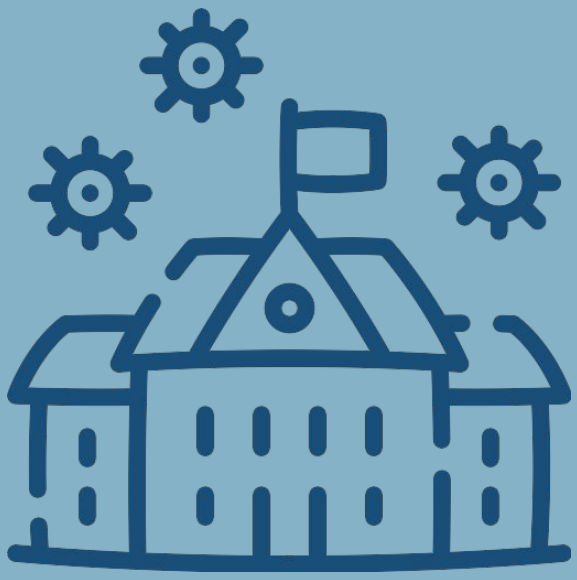
CONTENTS

01 Problem Statement

Background

In 2020, the COVID-19 pandemic affected the education ecosystem, impacting students, teachers, and school districts. In order to continue melding young minds, school districts deployed online learning tools to facilitate classes during quarantine.

In light of this historic moment, we as a data science team have information at our fingertips to discover how school districts handled the pandemic and if their efforts to deploy online tools were truly successful



Objectives

1. **To explore the data available to us. This includes but is not limited to: identifying frequently used learning platforms, identifying demographics that may have had less access to online learning, and identifying geographic regions that may have had less access to online learning**

2. **To quantify differences in tech engagement across different populations. This will be accomplished using various statistical models.**

3. **To provide educators with resources to inform them of tech engagement level in their district and top ed-tech products. Educators can leverage the dashboard and chat bot that our team builds to advocate for more resources.**

For future directions, school districts may use this report and the deliverables we create to advocate for better funding in EdTech online solutions. Ensuring that every student has the tools they need to succeed is our objective.

01 Project Goals

Goal 1
Exploratory Data Analysis

Through EDA, our team hopes to implement visualizations to better understand the data and interactions between variables.

Goal 2
Model Building

Through Modeling, our team hopes to quantify the relative digital engagement across different demographics.

Goal 3
Deliverables

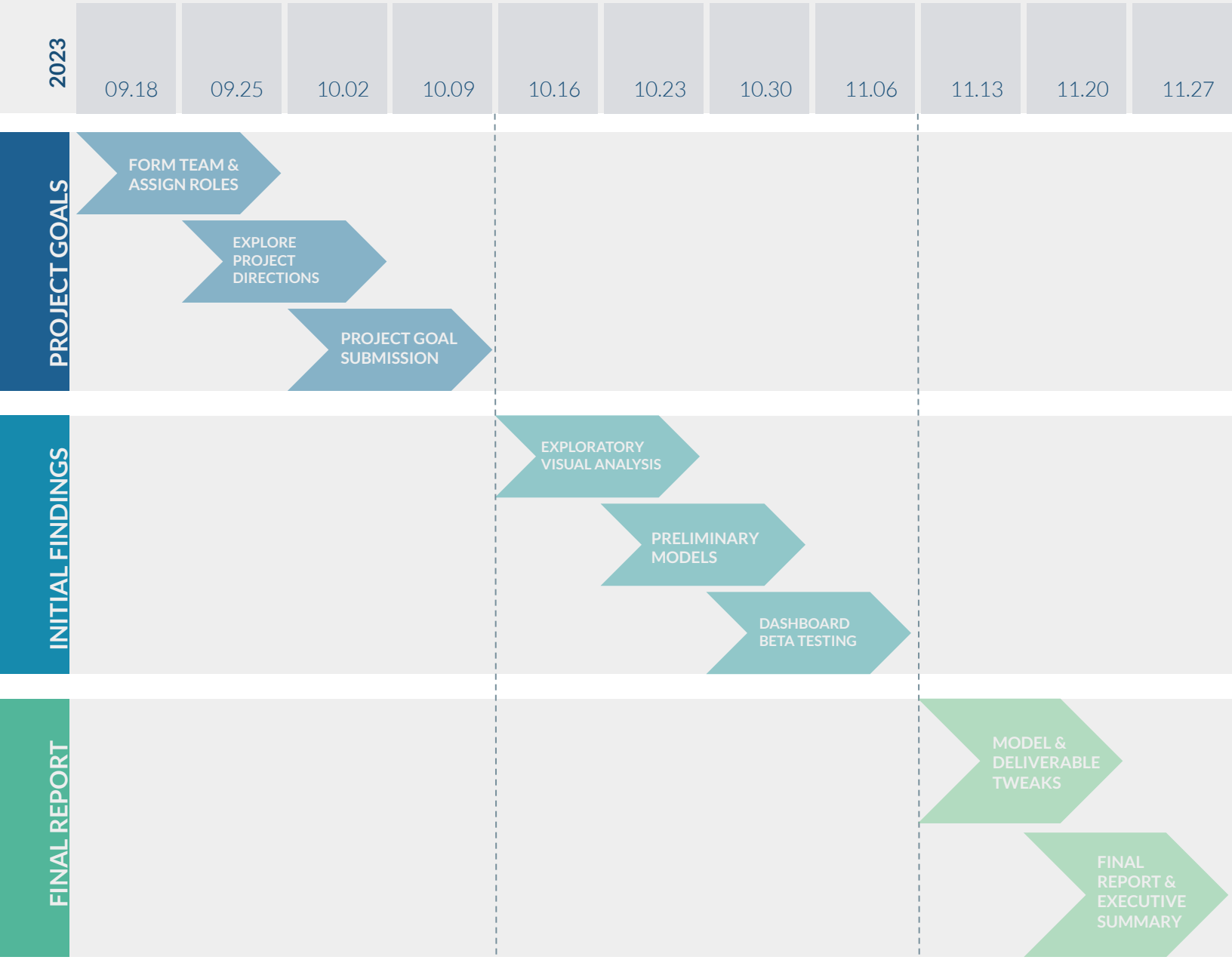
Through Deliverables, our team hopes to advise educators with a dashboard and chatbot.

*Cody Bakley, Julia Elliott, Mary Styers, Scott McQuiggan, Zarifa Zakaria. (2021). LearnPlatform COVID-19 Impact on Digital Learning. Kaggle. https://kaggle.com/competitions/learnplatform-covid19-impact-on-digital-learning

01 Project Charter

Our team conducted a 10-week project comprised of 3 separate sprints: setting up project goals, presenting initial findings to the CEO, and submitting a final report with an oral presentation. In the time since the initial findings, we have added the following:

1) Built 2 additional predictive models: decision tree and random forest, 2) went from beta testing our chat bot to a complete product, 3) added a geographical component to our dashboard and mobile app, 4) additional analysis write up on EDA visualizations.



What's Next? | A Deep Dive into the Dataset

In order to address our problem statement and complete our project goals, we must first take a look at the data we wish to utilize.

In the next few sections we will examine the datasets that our team used as well as any cleaning and transformation that was performed on the dataset.

Some key concepts that we will cover in these upcoming sections:

DATA

Data is our key resource, like raw material, and is the information we will use to understand and solve our problem statement We will use different types of information, like user engagement data and information retrieved from various digital learning platforms. The data we worked with was static, or collected during a set time frame.

VARIABLES

Variables are specific pieces of information that are crucial for our analysis. They are the categories in the dataset that help us understand the relationship between the data and real-life concepts and interpretations. Some variables we will look at are how often students engage with digital learning platforms and the student demographics of school districts.

BEST PRACTICES

Accurately collect data from various reliable sources for trustworthy results.

Handle data responsibly by respecting privacy and legal considerations.

Have transparency with the methods and purpose of data transformation and cleaning methods.

02 Description of the Data

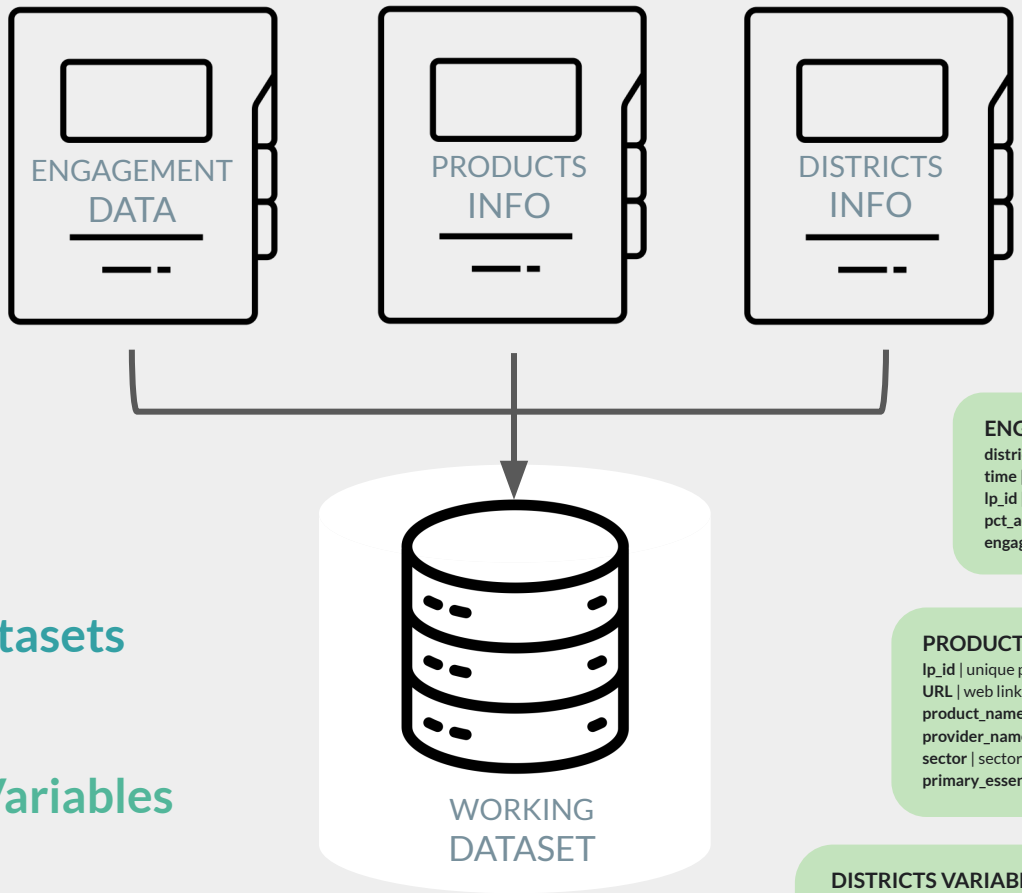
BACKGROUND

The datasets that are provided contain information from edtech engagement during 2020. Over 200 school districts provided daily engagement data, which was then combined with other external sources to create modified and constructed variables.

Engagements: This file provides information on the percentage of students who logged into digital platforms such as Google Docs, YouTube, Canvas, Schoology, and more to access educational content. It tracks their online activities, specifically the pages they loaded for learning or listening purposes.

Products: This file encompasses a list of all the digital products utilized by schools to deliver online education to students during the pandemic.

Districts: This file contains a list of district IDs across various US states where school districts employed digital platforms and tools for educational purposes.



ENGAGEMENT VARIABLES
district_id | district ID
time | date
lp_id | unique product ID
pct_access | % students 1-pg load/day
engagement_index | total pg load/1k students/day

PRODUCTS VARIABLES
lp_id | unique product ID
URL | web link to product
product_name | name of product
provider_name | name of provider
sector | sector of education
primary_essential_function | function of product

DISTRICTS VARIABLES
district_id | district ID
state | US state of district
locale | NCES locale classification of US territory
pct_black.hispanic | % students black/hispanic
pct_free.reduced | % students free/reduced lunch
county_connections_ratio | % high speed connection
pp_total_raw | per pupil total expenditure

3 Datasets

23 Variables

22+ Million Data Points

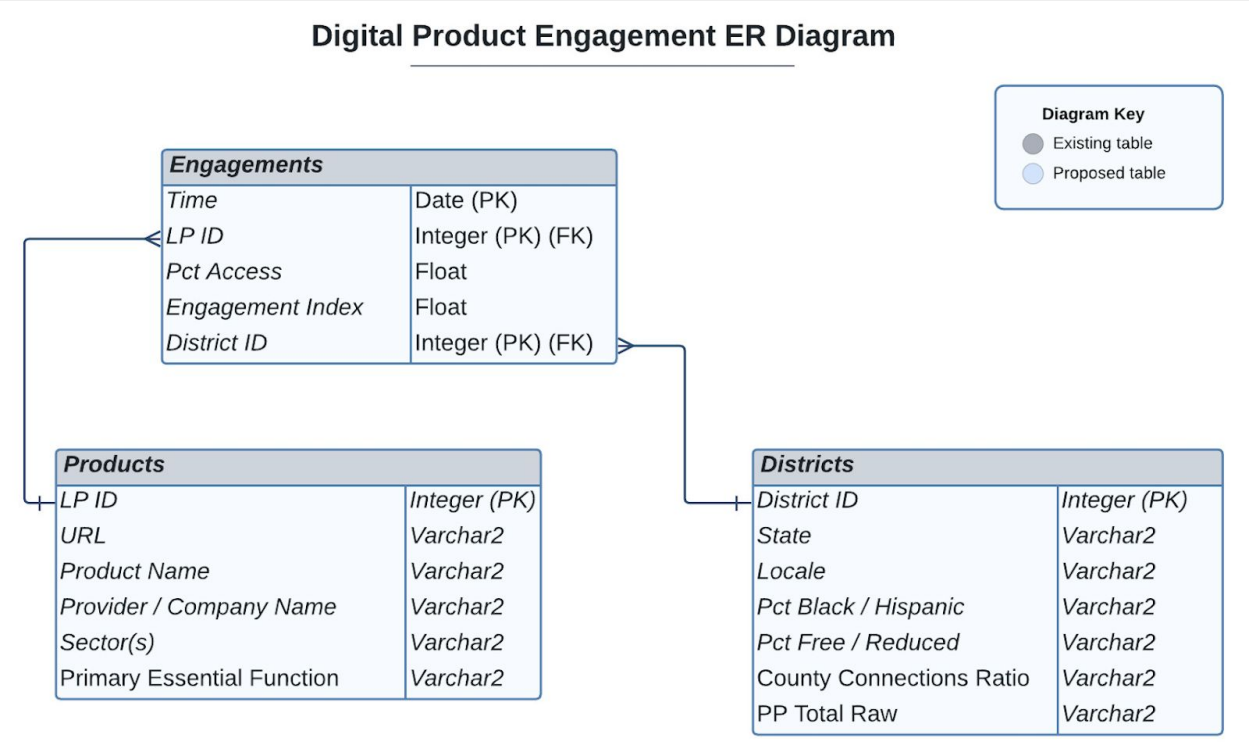
*Cody Bakley, Julia Elliott, Mary Styers, Scott McQuiggan, Zarifa Zakaria. (2021). LearnPlatform COVID-19 Impact on Digital Learning. Kaggle. https://kaggle.com/competitions/learnplatform-covid19-impact-on-digital-learning

03 Overview of the Data

Missing Values, Removed Variables, & Outliers

During this phase, we performed essential data cleaning to remove duplicates, handle empty columns or NaN values, and address outliers affecting the average engagement index and student access rate. Missing values were replaced with the mean or median where relevant, and records with minimal impact (less than 2-3% of the total) were removed.

Detailed Joining Methods



Concurrently, we established and validated relationships between tables from our data sources. The ER diagram below illustrates the data structure and elements involved. Learn Platform's software, deployed via Chrome extensions in 200 school districts, captured **two crucial elements** for our analysis in this dataset.

ENGAGEMENT INDEX

This metric represents the ratio of page-load events to the number of students, normalized per 1000 students.

PCT ACCESS

This metric indicates the number of students in a specific school district who accessed the digital platform on a given day.

04 Description of Transformation of Data

Variable Construction & Transformation

CONVERTED

The following numeric variables were loaded as character ranges such as "[a, b)." They were converted to lower bound integers for a more comprehensive analysis.

`Pct_black.hispanic`
`Pct_free.reduced`
`County_connections_ratio`
`pp_total_raw`

COLLAPSED

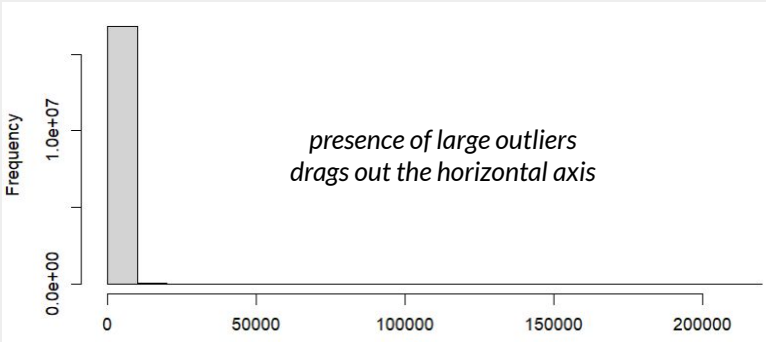
The following are categorical variables that were collapsed into less options for a more comprehensive analysis.

`Month_Yr`
`regions, pef2`
`Sector2`
`Semester`

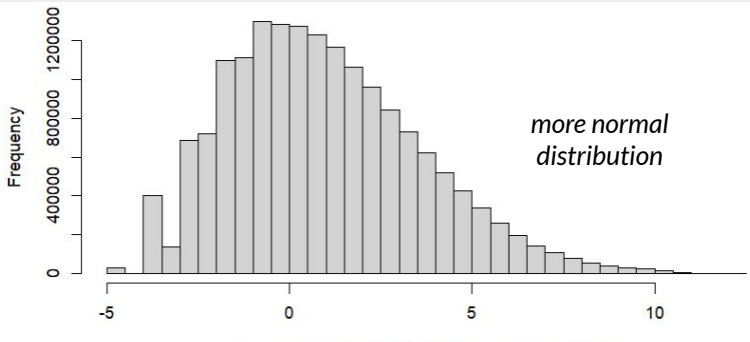
TRANSFORMED

`engagement_index` refers to "Total page-load events per one thousand students of a given product and on a given day." As an important outcome variable to our analyses, we performed a logarithmic transformation to yield a more normal distribution for further statistical exploration.
`log_engagement_index`

ENGAGEMENT INDEX
BEFORE LOG TRANSFORMATION



ENGAGEMENT INDEX
AFTER LOG TRANSFORMATION



VARIABLES WITHOUT TRANSFORMATION	MODIFIED VARIABLES
lp_id district_id time engagement_index state locale pct_access URL Product.Name Provider.Company.Name Sector.s. Primary.Essential.Function	**Variable Name : [structure] unique variable levels** pct_black.hispanic : num 0.2, 0.4, 0.6, 0.8, 1.0 (transformed) pct_free.reduced : num 0.2, 0.4, 0.6, 0.8, 1.0 (transformed) county_connections_ratio : num 1.0, 2.0 (transformed) pp_total_raw : num 6000, 8000, 10000, …, 34000 (transformed) log_engagement_index : num log(engagement_index) Month_Yr : chr format(as.Date(time), "%Y-%m") regions : chr "Northeast" "Midwest" "South" "West" (from State) pef2 : chr "LC" "CM" "SDO" (from Primary.Essential.Function) sector2 : chr "PreK-12 (w/ Higher Ed)" "PreK-12 (w/o Higher Ed)" (from Sector.s) Semester : chr "Winter_Spring" "Fall_Winter" "Summer" (from Month_Yr)

What's Next? | Using Our Data to Learn Something New

With a cleaned and transformed dataset, we can utilize our dataset to understand student engagement with online learning platforms. Our goal is to understand access to resources to set students up for success through digital learning.
In the next few sections, we will cover the Exploratory Data Analysis (EDA) which is multiple data visualizations. They will help us bridge the understanding between our dataset and our problem statement.

Some key concepts that we will cover in these upcoming sections:

CHARTS & GRAPHS

Charts and graphs are the visualizations that turn the data into pictures. From the charts we can understand how variables can relate to each other (which is called association) and how they might impact important variables. From the graphs, we can interpret the reasons why the data may behave in a certain way. The charts and graphs here were created in RStudio.

SCOPE & FRAMING

The way charts are presented can help guide you. Parts of the graph might be highlighted or emphasized to draw your eyes towards it. Other sections might be placed next to each other so they are more easily compared. Each of the 3 types of visualizations that are presented were carefully curated to frame the data in a way that is understandable.

BEST PRACTICES

First understand the purpose of the graph. Why are we looking at this data and what questions are we trying to answer?

Next, understand how to read the graph. Some graphs can be complex if you've never seen them before.

Lastly, interpret they key findings and takeaway overarching ideas about how the data interacts.

05 Analysis of the Data | EDA

PURPOSE

We are interested in examining student engagement with digital platforms over time because it is an indicator of access to resources that are essential for student success.

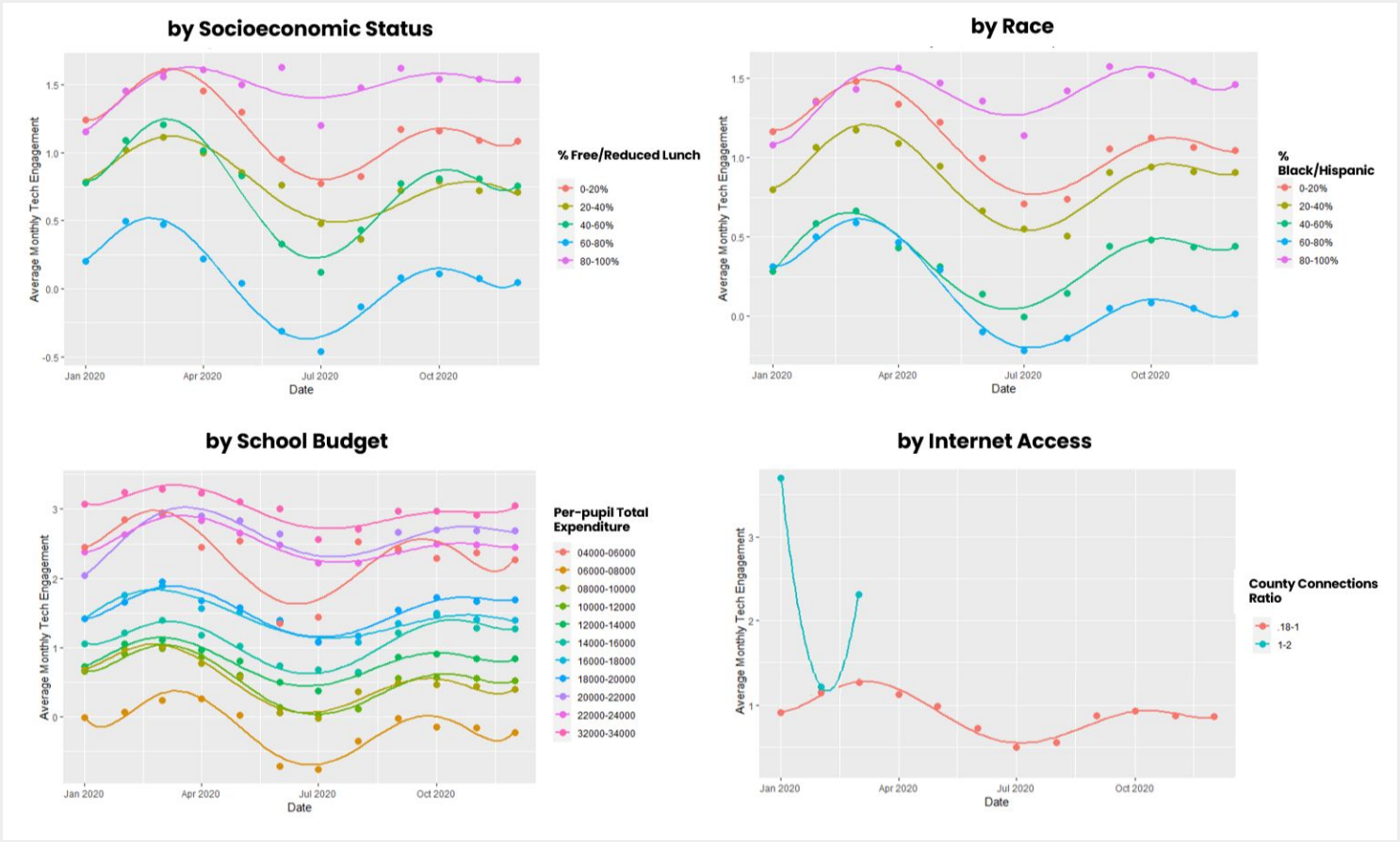
Different demographics may have different levels of tech engagement. Understanding those differences can help inform us on which groups may be at risk for lower engagement, and may need more resources.

VISUALIZATION

We are going to use a line chart on time-series data to look at trends during 2020 in . These line charts can be interpreted by reading left to right which indicates the passage of time.

Each line represents a the rising and falling student tech engagement of a specific demographic level. Demographics include socioeconomic status indicated by % of students that have free/reduced lunch, race indicated by % of students who are black/hispanic, school budget indicated by per-pupil total expenditure, and internet access indicated by county connections ratio.

STUDENT TECH ENGAGEMENT DURING COVID-19



INTERPRET

From the time-series line charts, we that see across all demographics, user engagement saw an increase in March during school closures and in September upon school reopening. As expected, there was a decrease over the summer months due to school being out of session.

Schools with higher expenditures for their pupils also have higher rates of average monthly tech engagement from their students as indicated by the pink and purple lines in the School Budget chart.

School districts with the 60%-80% of free/reduced lunch students have the lowest rates of average monthly tech engagement indicated by the blue line in the Socioeconomic Status chart.

Our dataset contains a subset of 3 months of data on internet access.

TAKEAWAYS

We can use this information in the following way:

From the time trends, we see an association between time and student tech engagement. These findings prompted the creation of a new time variable, 'Semester,' dividing the year into Fall/Winter, Winter/Spring, and Summer semesters. This enhances the interpretability of tech engagement analysis and provides a more suitable format for our statistical models moving forward.

From socioeconomic status, race, and school budget, we also see trends in engagement based off demographics that may cue us to their importance and influence on different populations. Including these variables when analyzing tech engagement will give us more accurate predictions.

05 Analysis of the Data | EDA

PURPOSE

We are interested in categorizing school district profiles because it will help signal if similar groups may be at risk for low engagement. The goal is to identify patterns or characteristics among school districts that could suggest challenges or barriers to technology engagement, helping to target interventions or support where needed.

VISUALIZATION

We are going to use a radar to look at demographic profiles. A radar chart is interpreted by the following: each color corresponds to different profile (e.g. different State). Each axis provides information about the profile being represented. Distance from the center indicates relative level of the corresponding attribute; lower relative levels are close to the center vs higher relatives are levels further out. Refer to the blue bordered profiles and how they compare on each axis to the other profiles.

DIGITAL LEARNING DEMOGRAPHIC PROFILES

INTERPRET

The distance from the center to a point on each spoke indicates the magnitude of that variable, and the overall shape of the chart illustrates patterns and relationships among the variables. These charts are useful for comparing the overall profiles of entities or observations across multiple dimensions.

The blue bordered shapes are the profiles we would like to focus on. They are Illinois (State), City (Locale), PreK-12 (Sector), Digital Learning (Function). Relative to the other profiles (shapes in their graph), these profiles have

- Relatively low rates of racial diversity (top axis is relatively close to center)
- Relatively low rates of reduced-lunch eligible students (left axis is relatively close to center)
- Relatively high or average access to reliable internet sources (right axis is relatively far from center)

TAKEAWAYS

We can use this information in the following way:

Because we have so many profiles that behave similarly, we can guess that tech engagement may be lower in schools with less access to the internet, and in school districts with less diversity and affluence. These populations should be brought into the conversation when it comes to funding access to online learning resources.

These demographic profiles (Illinois , City , PreK-12 , and Digital Learning Function), with predominantly white and affluent communities have better access to the internet. Based off these findings, these profiles do not seem to be at risk of low-tech engagement. We may use this as a basis for hypotheses in our models.

05 Analysis of the Data | EDA

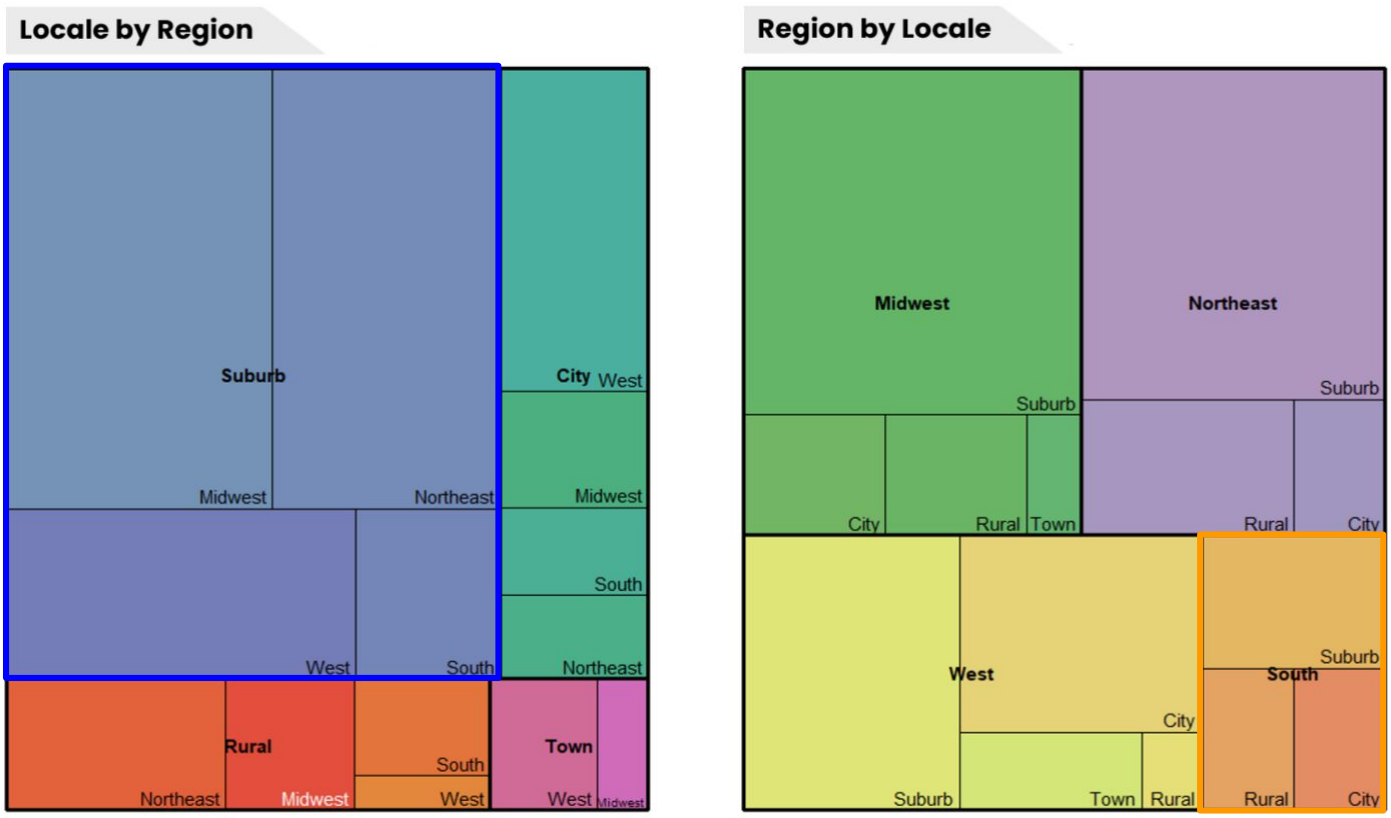
PURPOSE

We are interested in examining the distribution of data across different geographic areas because it will ensure a balanced or comprehensive representation for meaningful analysis and help us focus on the predominant characteristics of the populations we are analyzing.

VISUALIZATION

We are going to use a treemap to look at regional makeup of our dataset. A treemap is interpreted by the following: the more area a section color takes up, the more representation we have in our dataset. A smaller section would correspond to relatively less representation in the dataset. A section can be broken up into even smaller sections based on multiple variables. We chose this hierarchical chart to examine parts of the whole.

DIGITAL LEARNING DEMOGRAPHIC PROFILES



INTERPRET

From the 2 tree maps, we can compare relative differences across Locale in the United States (Suburb, City, Rural, Town) and Region in the United States (Midwest, Northeast, West, South).

The majority of the dataset is comprised by suburb data (outlined in blue) indicated by the largest section in the left treemap. It also shows that the South (outlined in orange) is underrepresented in this dataset indicated by the smallest section in the right treemap.

TAKEAWAYS

We can use this information in the following way:

When creating our models we can use these as reference groups and understand the bias incorporated within our dataset.

We want to use the groups with the most representation as our reference groups because we have the strongest grasp of how these geographic areas will behave, especially in terms of student tech engagement. With more statistical power, we will have a better ability to predict student tech engagement and better understand what populations need better access to resources.

What's Next? | Making Informed Decisions

From our EDA, we were able to discover the following:

LINE CHARTS

We identified time trends in our data that allowed us to create a semesters time reference which will enhance interpretability of our upcoming statistical models

Socioeconomic status, Race, and school budget are likely influencing the studied populations and need to be thought about with predictive.

RADAR CHARTS

We identified profiles that behave similarly and suggest a correlation between lower tech engagement and limited internet access, as well as lower diversity and affluence.

Many profiles focus on predominantly white and affluent communities. These profiles have better internet access which suggests they are not at risk of low tech-engagement.

TREE MAPS

We learned that selecting highly represented groups as reference points enhances our understanding of geographic areas' behavior in terms of student tech engagement, boosting statistical power for more accurate predictions and identifying populations in need of improved resource access. When creating our models we can use these as reference groups and understand the bias incorporated within our dataset.

It's crucial to include these underrepresented populations in discussions about funding access to online learning resources.

But now that we have hypotheses about different variables in our dataset, how can we quantify them?

In the upcoming sections, we'll explore Predictive Modeling, including diverse methods used to predict student tech engagement based on demographic variables. We'll compare model performances and discuss key findings.

Some key concepts that we will cover in these upcoming sections:

PREDICTIVE MODELS

Predictive models are like a machine. You give it our clean and transformed data to see how variables impact each other in order to answer questions and test hypotheses. The questions come from our problem statement, and the hypotheses come from our EDA. The models can tell us there are direct inverse relationships between variables, or maybe none at all.

PERFORMANCE & ERROR

How can we tell which model is the "best"? Well, it's the one that makes the least mistakes when predicting the outcome. We make the data using training dataset, and test it with separate data to see how accurately it can predict. The different models have different error rates, and the one with the smallest amount of error performs the "best."

06 Statistical Models

STATISTICAL MODELS

Statistical models take data, and predict an outcome based on patterns in the data. In our case, we looked at demographics of student school districts to explore their effects on engagement with digital learning platforms, and make informed decisions to stakeholders.

PREDICTORS & OUTCOME

There are 8 variables that we will consider when predicting engagement. The predictors are locale, regions, pef2, sector2, Semester, pct_black.hispanic, pct_free.reduced, pp_total_raw. The outcome is log_engagement_index: "total page-load events per one thousand students of a given product and on a given day." We want to know what variables impact engagement.

OUR WORK

There were 3 different types of models that we considered generalized linear regression, a decision tree, and a random forest. Of these types of models, the random forest yielded results with the lowest amount of error. We focus on the results from the random forest.

PREDICTOR	OPTIONS
locale	**"Suburb"** "City" "Rural" "Town"
regions (from state)	"Northeast" **"Midwest"** "South" "West"
pef2 (from Primary.Essential.Function)	**"LC"** "CM" "SDO"
Sector2 (from Sector.S)	**"PreK-12 (with Higher Ed)"** "PreK-12 (without Higher Ed)" "Other"
Semester (from Month_yr from time)	**"Winter_Spring"** "Fall_Winter" "Summer"
pct_black.hispanic	0.2, 0.4, 0.6, 0.8, 1.0
pct_free.reduced	0.2, 0.4, 0.6, 0.8, 1.0
pp_total_raw	6000, 8000, 10000, …, 34000

MODEL	ROOT MEAN SQUARE ERROR
generalized linear model (AIC- stepwise)	2.735586
decision tree (pruned)	2.848849
random forest (tuned)	**2.718192** (lowest error, highest accuracy)

06 Models | Generalized Linear Model (GLM)

A generalized linear model (GLM) is a statistical tool that helps you understand and model the relationship between variables contained within a dataset. It is efficient in making predictions and helping our team of data scientists draw conclusions about our data.

Although the GLM was not the best performing model, it does tell us some interesting patterns between the predictors in the model and our response: student engagement with online learning platforms:

GLM TAKEAWAYS

ENGAGEMENT BY LOCALE

Rural*** areas and Town** areas had significantly higher predicted engagement rates in comparison to Suburbs.

ENGAGEMENT BY REGION

South*** regions and West** regions had significantly lower predicted engagement rates in comparison to Midwest regions.

ENGAGEMENT BY SEMESTER

The Summer*** semester had significantly lower predicted engagement rates in comparison to the Winter/Spring semester.

The above mentioned were highly significant indicating a strong predictive capability (***p<0.001,**p<0.01, *p<0.1)

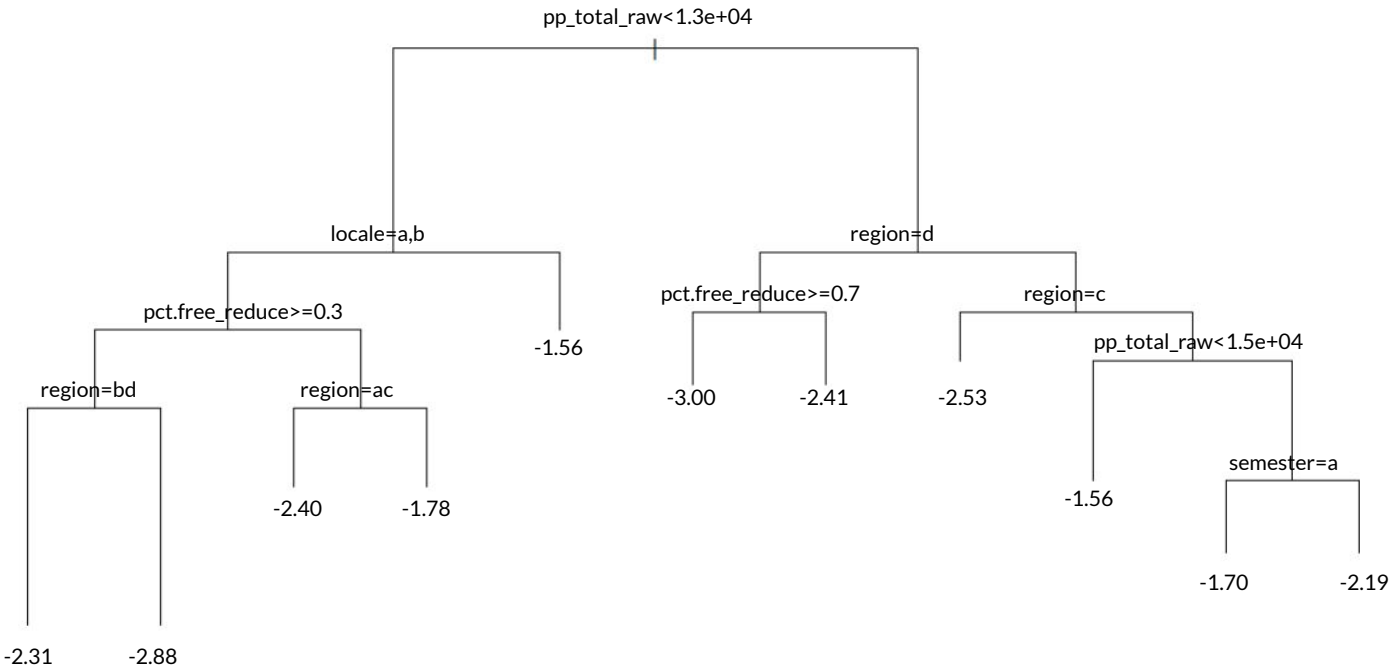
The next model that we tested for predicting engagement was the decision tree.

06 Models | Decision Tree

A decision tree functions as a dynamic roadmap guiding you through decision-making at each step. Comprising four key elements—root nodes, branches, decision nodes, and leaf nodes—it simplifies complex questions into manageable components.

The root node acts as the starting point, posing the initial question. Branches extend as diverse answers to this question. As you progress, decision nodes emerge, presenting new questions based on preceding insights. Culminating in leaf nodes, the decision tree unveils final outcomes, effectively breaking down intricate questions into digestible segments.

TREE CHART



DECISION TREE TAKEAWAYS

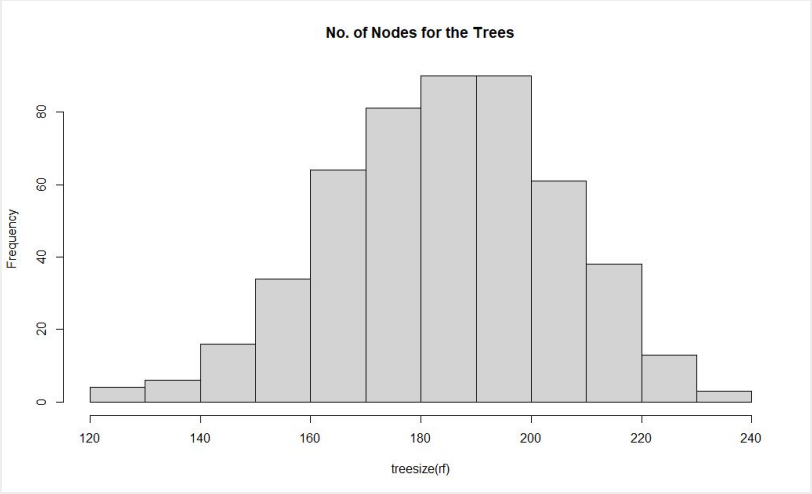
The decision tree we produced yielded the *poorest* results with the highest error. With the assistance of machine learning techniques, we constructed a random forest which is a product of aggregating results of many decision trees into one model. Our final model section will cover the results of our best performing model and some key takeaways

06 Models | Random Forest

Our team utilized an advanced **machine-learning** technique known as Random Forest. This sophisticated algorithm considers various factors by employing a random subset of data to create decision trees, effectively tackling both classification and regression challenges.

To enhance its effectiveness, we fine-tuned the algorithm by adjusting its parameters such as number of trees and nodes. This optimization process allows the Random Forest to make highly accurate predictions, ensuring its performance aligns seamlessly with our specific goals.

Our Random Forest Model had the highest level of accuracy across all models we ran and yielded the lowest level of error overall.

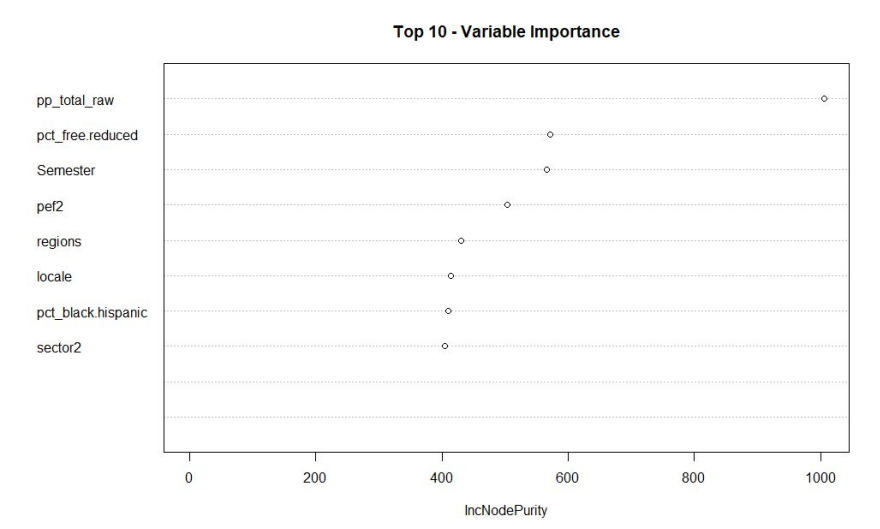


HYPERPARAMETER TUNING

Our final model utilized ~190 decision trees to aggregate result. This was the best fit to avoid overfitting and underfitting the model.

VARIABLES OF IMPORTANCE

Our model tells us that pp_total_raw, pct_free.reduced, and semester are the 3 variables of highest importance in predicting engagement from our dataset



06 Models | Random Forest

VARIABLES OF HIGH IMPORTANCE TAKEAWAYS

PER PUPIL TOTAL EXPENDITURES

School districts with lower per pupil total expenditures had significantly lower predicted engagement rates.

% FREE/REDUCED LUNCH

School districts with higher percentages of student eligible for free/reduced lunch had significantly low predicted engagement rates.

SEMESTER

School districts with during the summer had significantly lower predicted engagement rates than any other semesters.

FINAL MODEL TAKEAWAYS

Overall, we produced 3 separate statistical models to predict tech engagement from school district data in an attempt to discover what variables play a key role in the success of digital learning.

The generalized linear model performed well, and informed us that **Suburbs in the South and West regions had lower rates of Engagement.**

The random forest model performed best, and informed us that **school districts with lower per pupil total expenditures and high rates of free/reduced lunch had lower rates of Engagement.**

When consulting stakeholders, we must consider their regional makeup, the funding their school has available for students, and the affluence of the neighborhood. These variables play the biggest role in students' access to digital learning.

What's Next? | From Models to Solutions

Statistical and predictive models help us identify region and population with low edtech engagement. We then come up with practical solutions with cutting edge deliverables to address the gaps.

To inform stakeholders on **Who** needs the resources to improve edtech engagement, and **What** digital learning solutions can align with their needs, our team has created a **dashboard, a mobile dashboard, and a chatbot. Our goal is student success via digital learning, and these deliverables utilize the data and findings to help stakeholders make informed decisions.**

Some key concepts that we will cover in these upcoming sections:

DASHBOARD

Our Tableau Dashboard provides educators and parents with an immersive experience of interactive data modeling.

Users navigate effortlessly between student engagement analysis and digital learning platform insights, placing analytical power directly in the hands of users.

Complemented by a sleek mobile application, this dashboard is easily accessible on the go, ensuring a seamless and convenient experience.

CHATBOT

Our EdTech Product Chatbot deliverable will fulfill the mission of LearnPlatform by Instructure by addressing inquiries from educators about online learning solutions, enabling them to make well-informed choices.

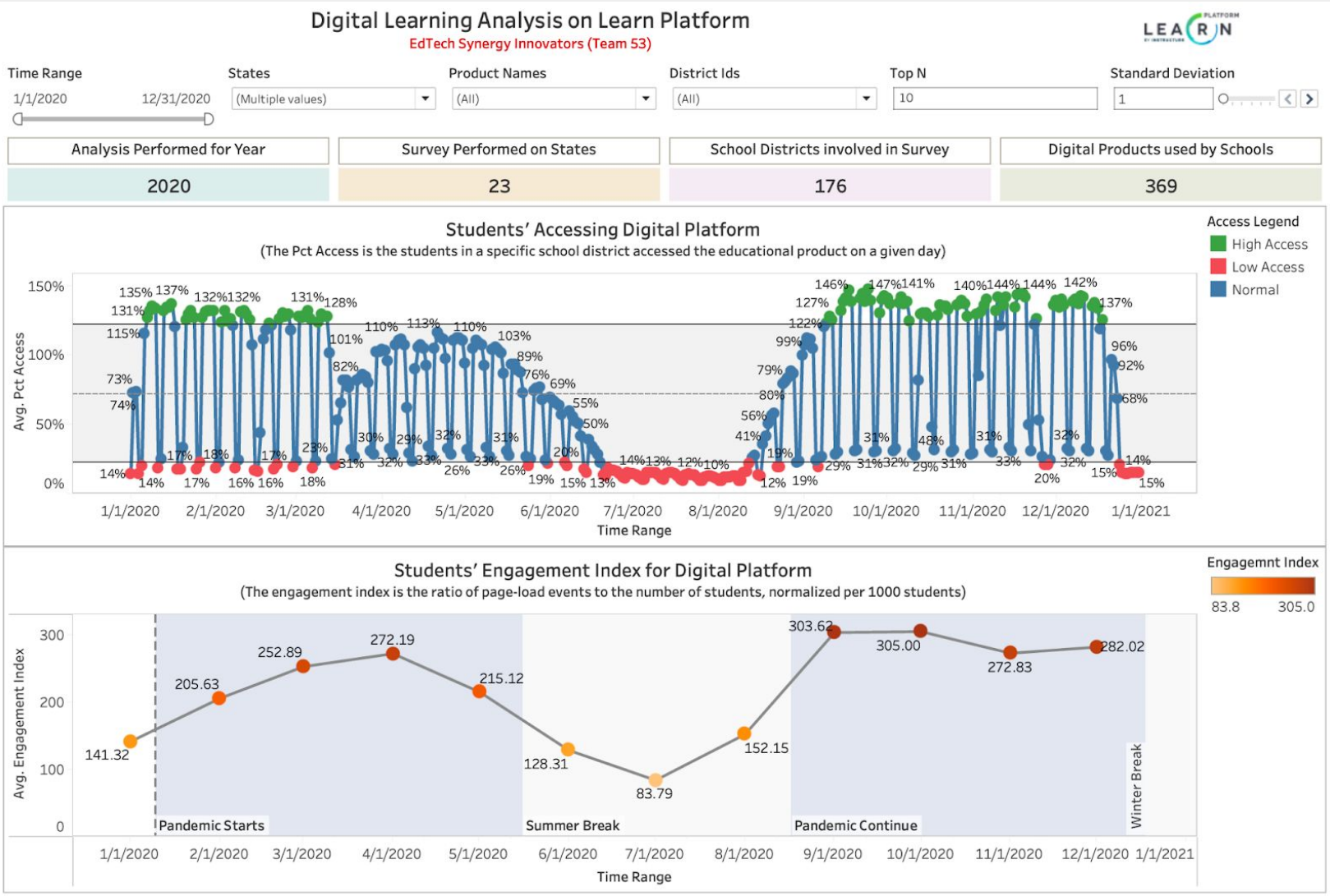
This chatbot is an artificial intelligence (AI) bot that employs the OpenAI's ChatGPT algorithms to generate new responses based on the ed-tech products information we gathered through their websites

The chatbot is capable of answering common educational platform inquiries and needs. In essence, our chatbot closes the loop, embodying a seamless transition from artificial intelligence to a user-friendly, practical solution.

07 Dashboard | Student Engagement

For the purpose of our analysis, our team began the development of a visualization dashboard to gain deeper insights into student profiles and the digital products most effectively used by students during the pandemic. We employed Tableau as our visualization tool to construct two primary components as shown below:

Student's Engagement Analysis



https://public.tableau.com/app/profile/rohan.tandon/viz/LPEngagementAnalysis/LPEngagementAnalysis

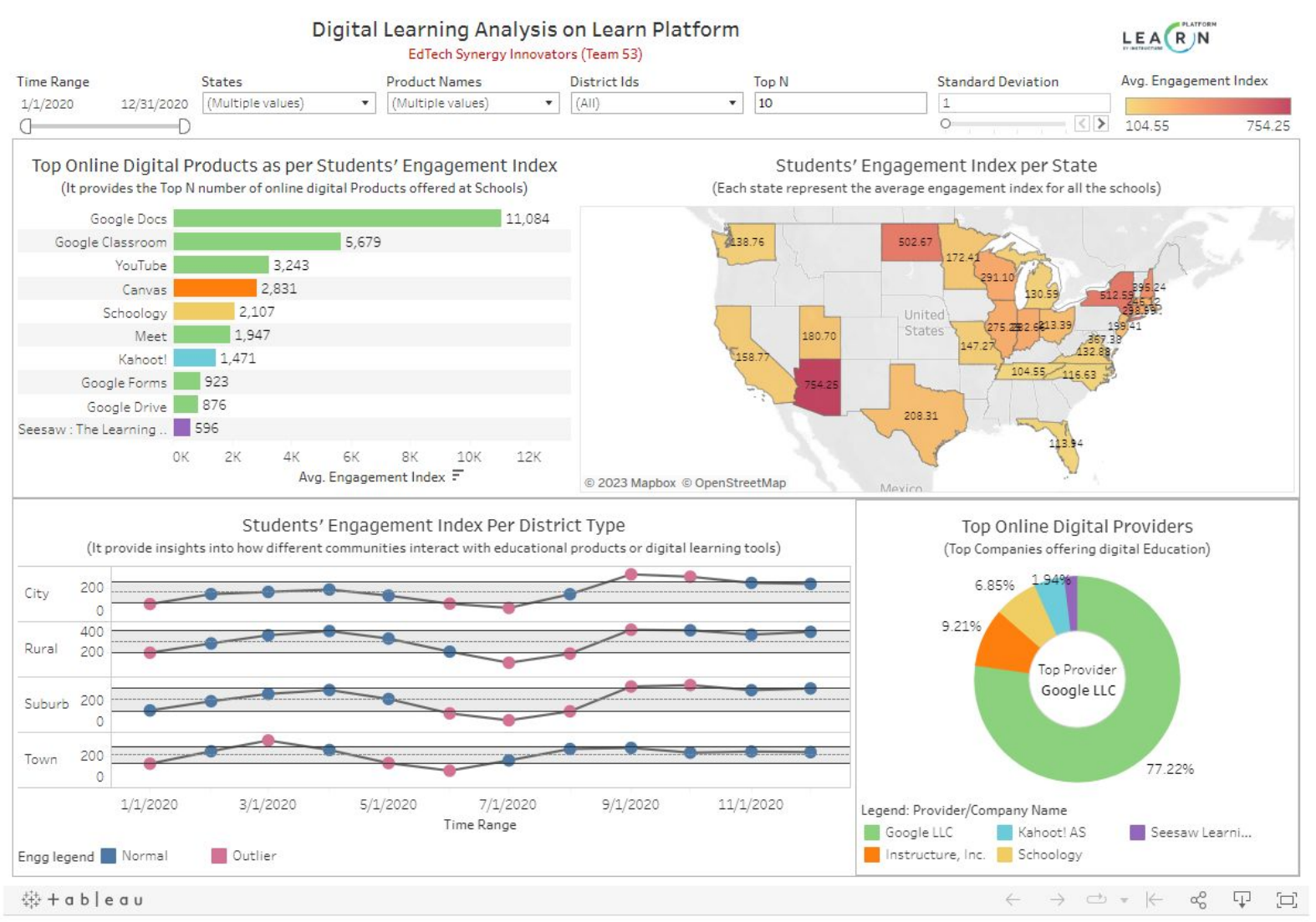
These components allowed us to understand the trends in online digital platform access by students before and after the pandemic's onset, covering the entirety of 2020, including summer and winter breaks. We also employed the standard deviation technique to identify outliers for various data elements, thereby identifying high and low engagement indexes.

Of utmost importance was our goal to understand the top online products used by school districts to deliver education to students, along with a breakdown of the engagement index at the district level, which encompassed city, rural, suburb, and town.

07 Dashboard | Product Analysis

EdTech Synergy Innovators analyzed data from Learn Platform across US school districts. Student engagement with online learning increased from Feb to May 2020, dipped during summer and winter breaks when schools closed, showing fluctuations in engagement over these periods.

Learn Platform Product Analysis



Among the top 10 products used by school districts, six were offered by Google LLC (Google Docs, Google Classroom, YouTube, Meet, Google Forms, Google Drive), followed by Instructure, Inc. (Canvas) and Schoology (Schoology).

Interestingly, we observed that there was still some level of engagement during summer and winter breaks, which indicated that schools offered online education to students who may not have had access to the online educational platform during the regular academic year due to internet shortage, digital products at home or missing knowledge around the digital product usage. Furthermore, our product analysis disclosed that Google LLC emerged as the top service provider, extensively utilized by schools to deliver online educational content to students and teachers during the pandemic.

07 Dashboard | Tableau Mobile

EdTech Synergy Innovators will leverage Tableau Mobile Dashboard, providing us with the capability to access all data and conduct future analyses conveniently through a mobile application at our fingertips.

Tableau Mobile

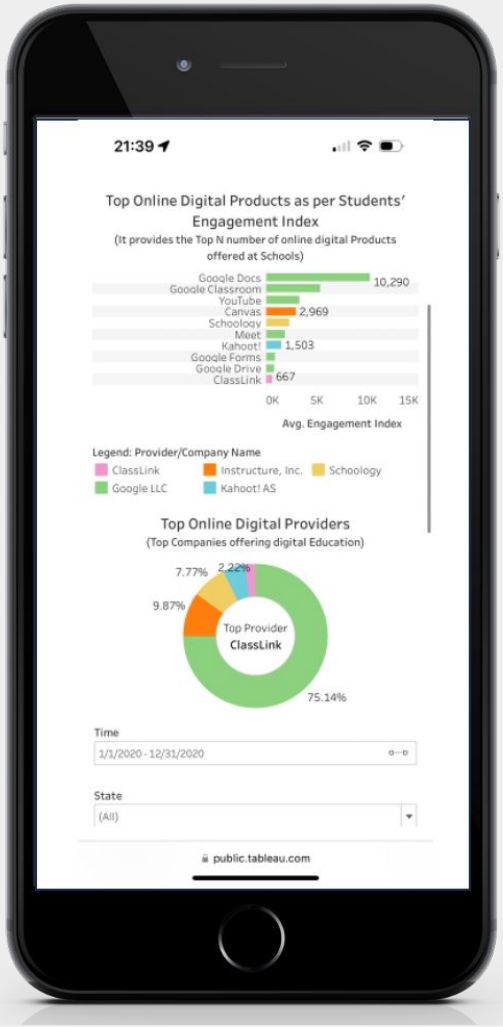
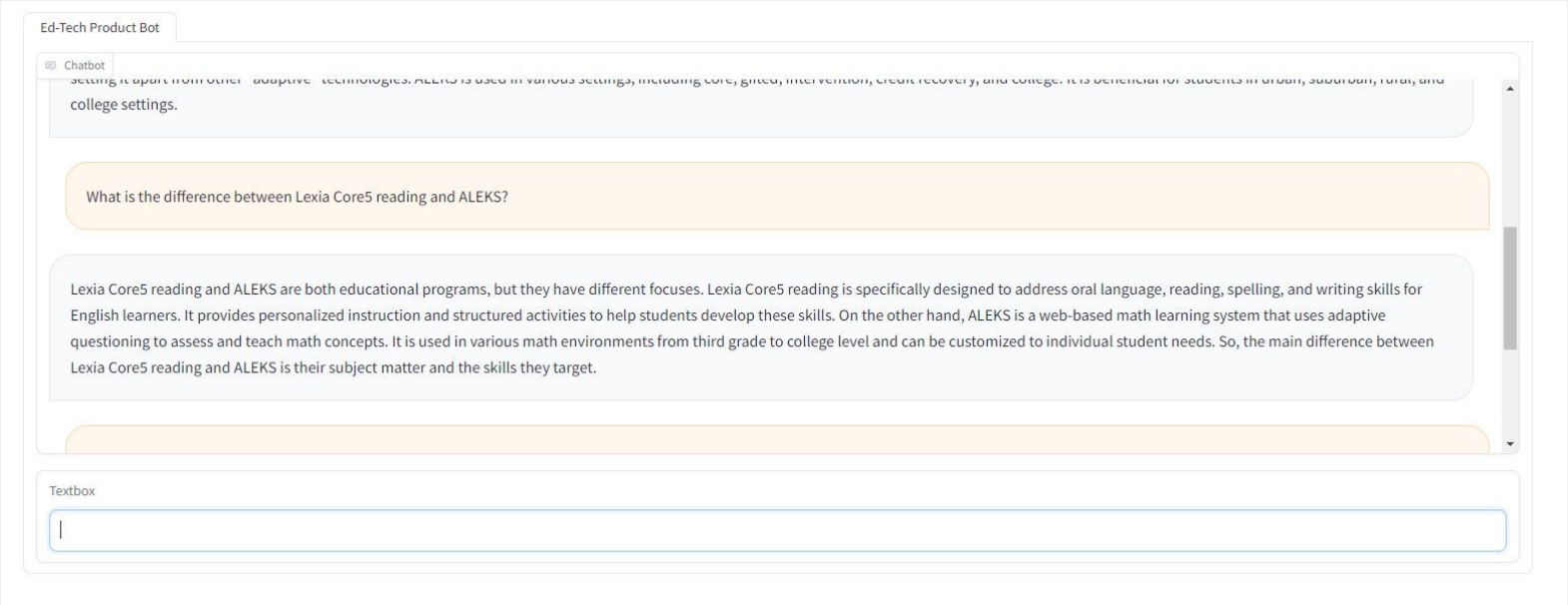


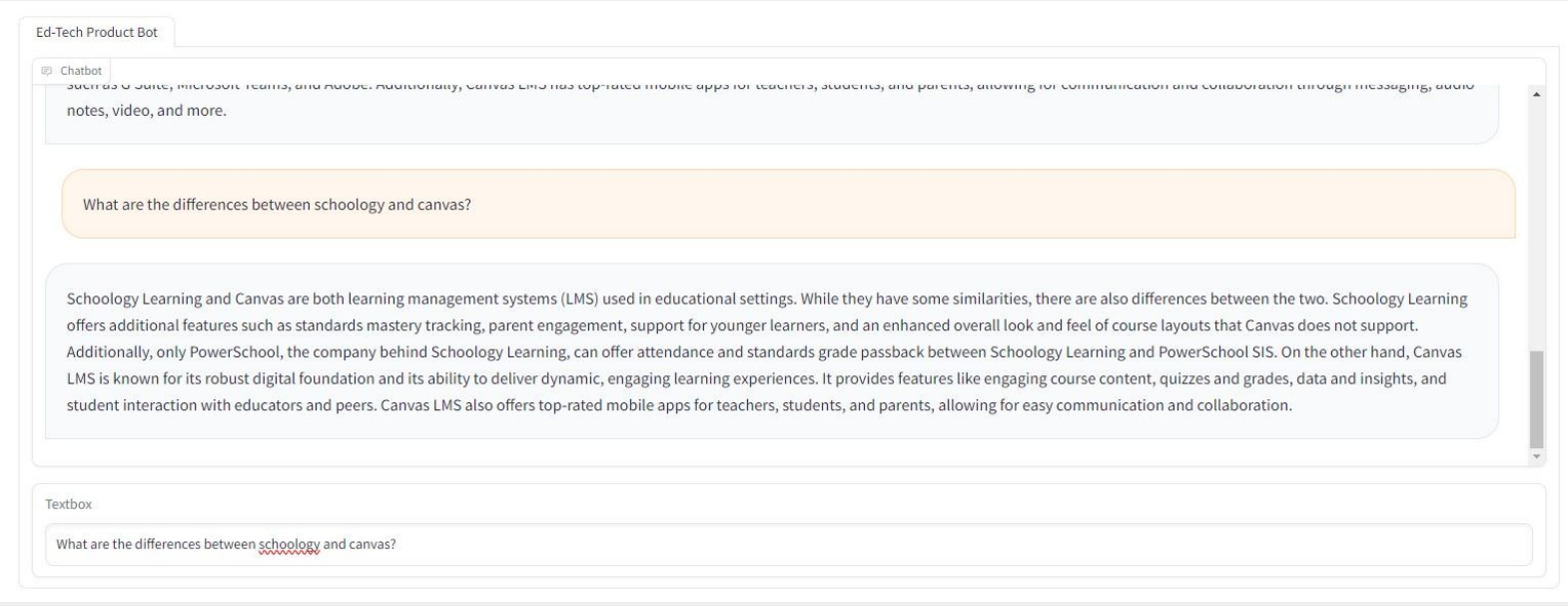
Tableau Mobile allows easy access to Tableau Server or Tableau Cloud content, enabling effortless exploration of data using a mobile device. It offers secure access to analytics, ensuring a user-friendly experience whether online or offline, on a tablet or smartphone.

08 Chatbot

An Ed-Tech Product Chatbot, powered by OpenAI's ChatGPT, has been developed to respond to inquiries about various online learning solutions. Integrated with a user-friendly web interface through Gradio, this chatbot can compare online learning solutions and assist educators in identifying products that best suit their requirements, especially in a landscape where there is an abundance of online offerings.

- **Customer Support and Engagement** answer common questions about prominent online learning solutions, offering assistance that enhances user satisfaction and engagement.
- **Data Analytics** provide valuable insights into user behavior, preferences, and pain points through Chatbot data collection.
- **Customization and Integration** tailor to the specific needs of LearnPlatform and integrate with existing systems and processes.





09 Conclusions & Recommendations

RECAP OF CONCLUSIONS

EXPLORATORY DATA ANALYSIS

Time series visualization identified trends, enabling a semesters time reference.

The radar charts revealed correlations between low tech engagement, limited internet access, and specific demographics.

The treemaps emphasized using highly represented groups as reference points for accurate predictions and mitigating bias in models.

MODELING

From these models, we were able to quantify the impact of the following different demographics on student tech engagement.

Suburban locale had lower rates of Engagement.

School districts with lower per pupil total expenditures and high rates of free/reduced lunch had lower rates of Engagement.

DELIVERABLES

The Tableau Dashboard is a user-friendly tool for exploring student engagement and learning data. With a mobile app, it's easy to use on the go.

Our EdTech Chatbot employs artificial intelligence to help with online learning platform questions, offering an efficient way for educators to pinpoint products that align with their specific needs.

RECOMMENDATIONS FROM OUR TEAM

Our goal from the beginning was to ensure that all students have access to digital learning resources to ensure their success in school no matter what demographics they come from.

- We recommend the continued research and collaboration of school districts with data scientists to identify at-risk populations of low tech engagement rates. Ensuring they have access to the resources is key to students' success. These populations may include school districts of Suburban locale, low per-pupil expenditures, high populations of free/reduced and Black/Hispanic student populations.

- We recommend the continued improvement of the deliverables. In predictive modeling, we covered 3 different models, but would recommend testing different models like XGBoost which is a machine-learning model that could handle the mix of categorical and numerical variables from the dataset. This type of model could provide higher predictive accuracy.
- In the dashboard, we recommend the incorporation of data filters when consulting different geographical regions and locales. Seeing the importance of geolocational variables from our EDA, tuning the dashboards for personalized consulting sessions with school districts will make the data you present representative of the communities you interact with.

- In the chat bot, we recommend making it accessible through different online portals to school administrators who may have questions about digital learning resources. Having that type of assistance at their fingertips will help them make informed decisions about providing the resources and funding for their students' success. Another aspect to consider is continued work on making the chatbot available in different languages. Many school districts do not have English spoken as their primary language, and they need access to digital learning resources just like any other school district.

Overall, there is still more work that can be done on the front of research, but we hope that our recommendations provide a solid launching point to advise school districts on the next steps they can follow to ensure the success of each and every student via digital learning resources.

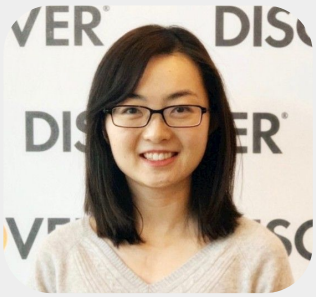
Appendix | Meet the Team

Kay Quiballo | Model Developer, Graphic Designer

Kay Quiballo is a data analytics professional with 4 years of experience in Alzheimer's and LGBT research. Kay has a bachelors in mathematics and statistics, and anticipates a masters in data science in December 2023. He specializes in topics such as analytical modeling, statistics, and machine learning.

Lauren Vaught | Program Manager, App Developer

Lauren Vaught is an operations research analyst and application developer. Lauren has over 12 years of experience in academic research with a bachelors of science in applied physiology and kinesiology with a minor in business. She anticipates receiving her masters in data science in December 2023 with an emphasis on analytics and predictive modeling as well as artificial intelligence.



Ge Li | Data Analyst, App Developer

Ge Li is a data science professional with over 8 years of experience in data analysis and quantitative research in the financial and epidemiology areas. Ge has a previous education background in psychology and school counseling. She anticipates receiving her masters in data science in December 2023 with an emphasis on analytics and modeling.

Rohan Tandon | Program Manager, BI Expert

Rohan Tandon is a proficient data expert in business intelligence, data engineering, and data analytics. With over 14 years of experience in Financial Investment Banking as a Subject Matter Expert (SME) in the Payments, Loans, and Trade Finance domains, he also possesses expertise in Retail Product Sales, Crediting/Commission, and Cloud-based Security Assurance engineering. He is currently pursuing an MS in data science with a major in Natural Language Processing and Artificial Intelligence, with an expected completion date in December 2023.

Brandon Eubank | Data Manager

Brandon Eubank has spent the last two decades in data driven manufacturing, financial services, and consulting. He is currently a startup founder and Partner with Growth Advisors investment bank. In his free time, he enjoys exploring the National Park Service and other public lands. Brandon has previous education experience in Economics and will receive a M.S. in Data science in December 2023.

Appendix | R Output & References

R Output | Structure of Dataset & Generalized Linear Model

```
> str(engagement_data_detailed2)
'data.frame':       22243691 obs. of  23 variables:

Variables with no transformation
 $ lp_id                     : num  10003 10003 10003 10003 10003 ...
 $ district_id               : int  3390 1052 1705 1791 5510 2074 2074 4668 8702 8815 ...
 $ time                      : Date, format: "2020-03-04" "2020-11-19" "2020-05-18" …
 $ engagement_index          : num  0.15 NA NA NA NA 0.07 0.14 0.02 NA NA …
 $ state                     : chr  NA "Illinois" "Washington" "Virginia" ...
 $ locale                    : chr  NA "Suburb" "City" "City" ...
 $ pct_access                : num  0.02 0 0 0 0 0.01 0.01 0 0 0 ...
 $ URL                       : chr  NA NA NA NA ...
 $ Product.Name              : chr  NA NA NA NA ...
 $ Provider.Company.Name     : chr  NA NA NA NA ...
 $ Sector.s.                 : chr  NA NA NA NA ...
 $ Primary.Essential.Function: chr  NA NA NA NA ...

Modified Variables
 $ pct_black.hispanic        : num  0.2, 0.4, 0.6, 0.8, 1.0 (transformed)
 $ pct_free.reduced          : num  0.2, 0.4, 0.6, 0.8, 1.0 (transformed)
 $ county_connections_ratio  : num  1.0, 2.0 (transformed)
 $ pp_total_raw              : num  6000, 8000, 10000, …, 34000 (transformed)
 $ log_engagement_index      : num  log(engagement_index)
 $ Month_Yr                  : chr  format(as.Date(time), "%Y-%m")
 $ regions                   : chr  "Northeast" "Midwest" "South" "West" (from State)
 $ pef2                      : chr  "LC" "CM" "SDO" (from Primary.Essential.Function)
 $ sector2                   : chr  "PreK-12 (with Higher Ed)" "PreK-12 (without Higher Ed)" (from Sector.s)
 $ Semester                  : chr  "Winter_Spring" "Fall_Winter" "Summer" (from Month_Yr)

> summary(glm1)

Call:
glm(formula = log_engagement_index ~ relevel(as.factor(locale),
    ref = "Suburb") + relevel(as.factor(regions), ref = "Midwest") +
    relevel(as.factor(pef2), ref = "Learning & Curriculum") +
    relevel(as.factor(sector2), ref = "PreK-12 (with Higher Ed)") +
    relevel(as.factor(Semester), ref = "Winter_Spring") + pct_black.hispanic +
    pct_free.reduced + pp_total_raw, family = "gaussian", data = engagement_data_detailed2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.5551  -1.9266  -0.2187   1.7126  10.9745

Coefficients:
                                                                                    Estimate    Std. Error  t value             Pr(>|t|)
(Intercept)                                                                       1.8647286485  0.0073644955  253.205 < 0.0000000000000002 ***
relevel(as.factor(locale), ref = "Suburb")City                                   -0.0233755807  0.0040569967   -5.762        0.00000000832 ***
relevel(as.factor(locale), ref = "Suburb")Rural                                   0.6703917292  0.0043780135  153.127 < 0.0000000000000002 ***
relevel(as.factor(locale), ref = "Suburb")Town                                    0.5062408580  0.0056487397   89.620 < 0.0000000000000002 ***
relevel(as.factor(regions), ref = "Midwest")Northeast                             0.2821204871  0.0071952106   39.209 < 0.0000000000000002 ***
relevel(as.factor(regions), ref = "Midwest")South                                -0.7165171157  0.0043713062 -163.914 < 0.0000000000000002 ***
relevel(as.factor(regions), ref = "Midwest")West                                 -0.3105787702  0.0033627057  -92.360 < 0.0000000000000002 ***
relevel(as.factor(pef2), ref = "Learning & Curriculum")Classroom Management      -0.2259996899  0.0053965119  -41.879 < 0.0000000000000002 ***
relevel(as.factor(pef2), ref = "Learning & Curriculum")School & District Operations  0.3939597282  0.0056345202   69.919 < 0.0000000000000002 ***
relevel(as.factor(sector2), ref = "PreK-12 (with Higher Ed)")Other               -0.6040098209  0.0151287422  -39.925 < 0.0000000000000002 ***
relevel(as.factor(sector2), ref = "PreK-12 (with Higher Ed)")PreK-12 (without Higher Ed) -0.0005832898  0.0026950190   -0.216                0.829
relevel(as.factor(Semester), ref = "Winter_Spring")Fall_Winter                   -0.0785412755  0.0028799881  -27.271 < 0.0000000000000002 ***
relevel(as.factor(Semester), ref = "Winter_Spring")Summer                        -0.8839798404  0.0039710814 -222.604 < 0.0000000000000002 ***
pct_black.hispanic                                                                0.0922123752  0.0102379699    9.007 < 0.0000000000000002 ***
pct_free.reduced                                                                 -0.7010269060  0.0105010027  -66.758 < 0.0000000000000002 ***
pp_total_raw                                                                      0.0000252670  0.0000004964   50.902 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 6.662361)

    Null deviance: 26590610  on 3808267  degrees of freedom
Residual deviance: 25371951  on 3808252  degrees of freedom
  (18435423 observations deleted due to missingness)
AIC: 18029696

Number of Fisher Scoring iterations: 2
```

Appendix | R Output & References

R Output | Stepwise Generalized Linear Model

```
NOTE: AIC STEPWISE yields the same model as the full glm


> step_car <- stepAIC(glm(log_engagement_index ~ ., data =
na.omit(inputData[,c("log_engagement_index","locale","regions","pef2","sector2","Semester","pct_black.hispanic","
pct_free.reduced","pp_total_raw")]),family="gaussian"), direction= "both")
Start:  AIC=12659.24
log_engagement_index ~ locale + regions + pef2 + sector2 + Semester +
    pct_black.hispanic + pct_free.reduced + pp_total_raw

                    Df Deviance   AIC
<none>                   17612 12659
- pct_black.hispanic  1    17626 12659
- pef2                2    17646 12660
- sector2             2    17648 12661
- pp_total_raw        1    17636 12661
- pct_free.reduced    1    17641 12662
- locale              3    17749 12674
- regions             3    17760 12676
- Semester            2    18026 12717


================================================================================

> summary(step_car)

Call:
glm(formula = log_engagement_index ~ locale + regions + pef2 +
    sector2 + Semester + pct_black.hispanic + pct_free.reduced +
    pp_total_raw, family = "gaussian", data = na.omit(inputData[,
    c("log_engagement_index", "locale", "regions", "pef2", "sector2",
        "Semester", "pct_black.hispanic", "pct_free.reduced",
        "pp_total_raw")]))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.2865  -1.9061  -0.2527   1.6926   9.3169

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.575e-02  6.436e-01   0.087  0.93098
localeRural                      6.941e-01  1.972e-01   3.519  0.00044 ***
localeSuburb                     8.688e-02  1.507e-01   0.577  0.56420
localeTown                       6.613e-01  2.237e-01   2.956  0.00315 **
regionsNortheast                 4.684e-01  2.832e-01   1.654  0.09832 .
regionsSouth                    -6.877e-01  1.659e-01  -4.145 3.51e-05 ***
regionsWest                     -3.681e-01  1.253e-01  -2.937  0.00334 **
pef2Learning & Curriculum        2.004e-01  2.000e-01   1.002  0.31661
pef2School & District Operations 5.918e-01  2.726e-01   2.171  0.03000 *
sector2PreK-12 (with Higher Ed)  1.210e+00  5.255e-01   2.302  0.02143 *
sector2PreK-12 (without Higher Ed) 1.228e+00 5.273e-01   2.329  0.01991 *
SemesterSummer                  -1.089e+00  1.519e-01  -7.171 9.62e-13 ***
SemesterWinter_Spring            4.083e-02  1.086e-01   0.376  0.70701
pct_black.hispanic               5.548e-01  3.795e-01   1.462  0.14394
pct_free.reduced                -8.173e-01  3.903e-01  -2.094  0.03637 *
pp_total_raw                     3.507e-05  1.838e-05   1.908  0.05647 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


--------------------------------------------------------------------------------

RMSE
> sqrt( mean( (testData[!is.na(testData$log_engagement_index),]$log_engagement_index-predict(step_car,
testData[!is.na(testData$log_engagement_index),]))^2 , na.rm = TRUE ) )
[1] 2.735586

(Dispersion parameter for gaussian family taken to be 6.628472)

    Null deviance: 18694  on 2672  degrees of freedom
Residual deviance: 17612  on 2657  degrees of freedom
AIC: 12659

Number of Fisher Scoring iterations: 2
```

Appendix | R Output & References

R Output | Decision Tree

```
> summary(dtree)
Call:
rpart(formula = formula, data = inputData, method = "class",
    control = rpart.control(minsplit = 30, cp = 0.001))
  n=11853 (3717 observations deleted due to missingness)

           CP nsplit rel error    xerror        xstd
1 0.002916702      0 1.0000000 1.0000000 0.001191024
2 0.001772898      1 0.9970833 0.9974264 0.001277283
3 0.001115210      4 0.9917646 0.9941666 0.001378237
4 0.001086615      5 0.9906494 0.9906494 0.001478776
5 0.001029424      8 0.9873896 0.9906494 0.001478776
6 0.001000000     10 0.9853307 0.9898773 0.001499850

Variable importance
       relevel(as.factor(regions), ref = "Midwest")                                     pp_total_raw
                                                 36                                               20
                                    pct_free.reduced          relevel(as.factor(locale), ref = "Suburb")
                                                 19                                               17
relevel(as.factor(Semester), ref = "Winter_Spring")                                  pct_black.hispanic
                                                  5                                                4

Node number 1: 11853 observations,    complexity param=0.002916702
  predicted class=-2.99573227355399  expected loss=0.9834641  P(node) =1



================================================================

> printcp(dtree)

Classification tree:
rpart(formula = formula, data = inputData, method = "class",
    control = rpart.control(minsplit = 30, cp = 0.001))

Variables actually used in tree construction:
[1] pct_free.reduced                                pp_total_raw
[3] relevel(as.factor(locale), ref = "Suburb")      relevel(as.factor(regions), ref = "Midwest")
[5] relevel(as.factor(Semester), ref = "Winter_Spring")

Root node error: 11657/11853 = 0.98346

n=11853 (3717 observations deleted due to missingness)

         CP nsplit rel error  xerror      xstd
1 0.0029167      0   1.00000 1.00000 0.0011910
2 0.0017729      1   0.99708 0.99743 0.0012773
3 0.0011152      4   0.99176 0.99417 0.0013782
4 0.0010866      5   0.99065 0.99065 0.0014788
5 0.0010294      8   0.98739 0.99065 0.0014788
6 0.0010000     10   0.98533 0.98988 0.0014999

================================================================ PRUNED
> summary(pdtree)
Call:
rpart(formula = formula, data = inputData, method = "class",
    control = rpart.control(minsplit = 30, cp = 0.001))
  n=11853 (3717 observations deleted due to missingness)

           CP nsplit rel error    xerror        xstd
1 0.002916702      0 1.0000000 1.0000000 0.001191024
2 0.001772898      1 0.9970833 0.9974264 0.001277283
3 0.001115210      4 0.9917646 0.9941666 0.001378237
4 0.001086615      5 0.9906494 0.9906494 0.001478776
5 0.001029424      8 0.9873896 0.9906494 0.001478776
6 0.001000000     10 0.9853307 0.9898773 0.001499850

Variable importance
       relevel(as.factor(regions), ref = "Midwest")
                                                 36
                                      pp_total_raw
                                                 20
                                    pct_free.reduced
                                                 19
         relevel(as.factor(locale), ref = "Suburb")
                                                 17
relevel(as.factor(Semester), ref = "Winter_Spring")
                                                  5
                                  pct_black.hispanic
                                                  4

Node number 1: 11853 observations,    complexity param=0.002916702
  predicted class=-2.99573227355399  expected loss=0.9834641  P(node) =1

================================================================

> printcp(pdtree)

Classification tree:
rpart(formula = formula, data = inputData, method = "class",
    control = rpart.control(minsplit = 30, cp = 0.001))

Variables actually used in tree construction:
[1] pct_free.reduced                                pp_total_raw
[3] relevel(as.factor(locale), ref = "Suburb")      relevel(as.factor(regions), ref = "Midwest")
[5] relevel(as.factor(Semester), ref = "Winter_Spring")

Root node error: 11657/11853 = 0.98346

n=11853 (3717 observations deleted due to missingness)

         CP nsplit rel error  xerror      xstd
1 0.0029167      0   1.00000 1.00000 0.0011910
2 0.0017729      1   0.99708 0.99743 0.0012773
3 0.0011152      4   0.99176 0.99417 0.0013782
4 0.0010866      5   0.99065 0.99065 0.0014788
5 0.0010294      8   0.98739 0.99065 0.0014788
6 0.0010000     10   0.98533 0.98988 0.0014999



================================================================

#Model Testing

#RMSE
rmse(testData[!is.na(testData$log_engagement_index),]$log_engagement_index, predict(pdtree, testData[!is.na(testData$log_engagement_index),]))
#returns 2.848849
```

Appendix | R Output & References

R Output | Random Forest

```
> rf <- randomForest(log_engagement_index~., data=
+                      na.omit(inputData[,c("log_engagement_index","locale","regions","pef2",
+                                           "sector2","Semester","pct_black.hispanic",
+                                           "pct_free.reduced","pp_total_raw")]), proximity=TRUE)
> print(rf)

Call:
 randomForest(formula = log_engagement_index ~ ., data = na.omit(inputData[,      c("log_engagement_index", "locale",
"regions", "pef2", "sector2",          "Semester", "pct_black.hispanic", "pct_free.reduced",          "pp_total_raw")]),
proximity = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 6.462547
                    % Var explained: 7.6

================================================================================

> train2 <- na.omit(inputData[,c("log_engagement_index","locale","regions","pef2",
+                                "sector2","Semester","pct_black.hispanic",
+                                "pct_free.reduced","pp_total_raw")])
> View(train2)
> t <- tuneRF(train2[,-1], train2[,1],
+             stepFactor = 0.5,
+             plot = TRUE,
+             ntreeTry = 150,
+             trace = TRUE,
+             improve = 0.05)
mtry = 2  OOB error = 6.50023
Searching left ...
mtry = 4         OOB error = 6.774974
-0.04226687 0.05
Searching right ...
mtry = 1         OOB error = 6.551135
-0.007831342 0.05

================================================
VARIABLE IMPORTANCE - reference the plot

> importance(rf)
                    IncNodePurity
locale                  414.2339
regions                 430.4845
pef2                    503.4345
sector2                 404.8261
Semester                566.0981
pct_black.hispanic      409.8437
pct_free.reduced        572.1560
pp_total_raw           1005.3107

================================================

#RMSE
sqrt( mean( (testData[!is.na(testData$log_engagement_index),]$log_engagement_index-predict(rf,
testData[!is.na(testData$log_engagement_index),]))^2 , na.rm = TRUE ) )
# returns 2.718192
```

References | Kaggle

Cody Bakley, Julia Elliott, Mary Styers, Scott McQuiggan, Zarifa Zakaria. (2021). LearnPlatform COVID-19 Impact on Digital Learning. Kaggle. https://kaggle.com/competitions/learnplatform-covid19-impact-on-digital-learning

Appendix | Color Scheme

DESIGN PALETTE

font: Lato

Color Palette (coolors)

D9ED92	B5E48C	99D98C	76C893	52B69A	34A0A4	168AAD	1A759F	1E6091	184E77
Mindaro	Light green	Light green	Emerald	Keppel	Verdigris	Bondi blue	Cerulean	Lapis Lazuli	Indigo dye

Logo

Banner

Icons:

Flaticon extension

Noun Project