Berkeley Engineering | BerkeleyHaas

PROFESSIONAL CERTIFICATE IN MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Office Hour #7 with Matilde D'Amelio

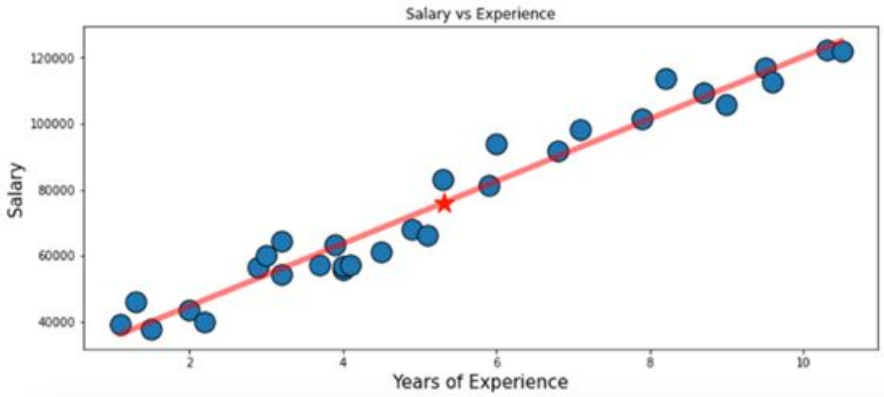
April 28, 2022 at 9 pm UTC

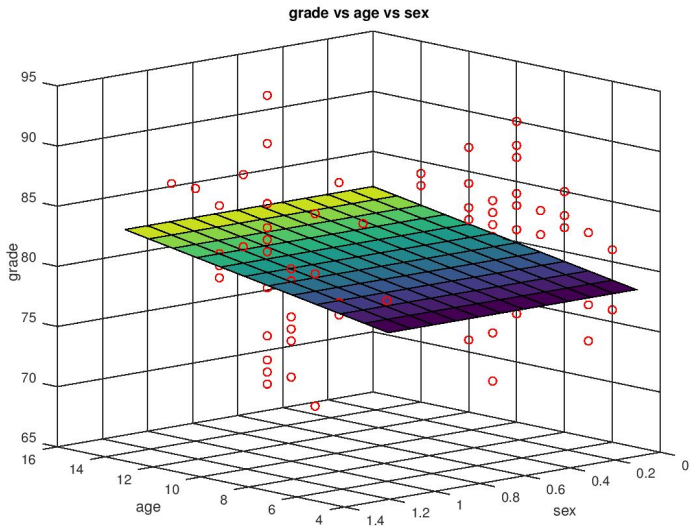
Linear Regression: Simple and Multiple

Identify the relationship between variables

1. Is there one variable (independent variable) that does a good job of predicting an outcome (dependent variable)?
2. Which variables are the best predictors of the outcome variable?
3. How much do these variables influence the outcome variable (indicated by the magnitude and sign of the beta estimates)?

$$y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + b$$





When do I use simple or multiple regression?

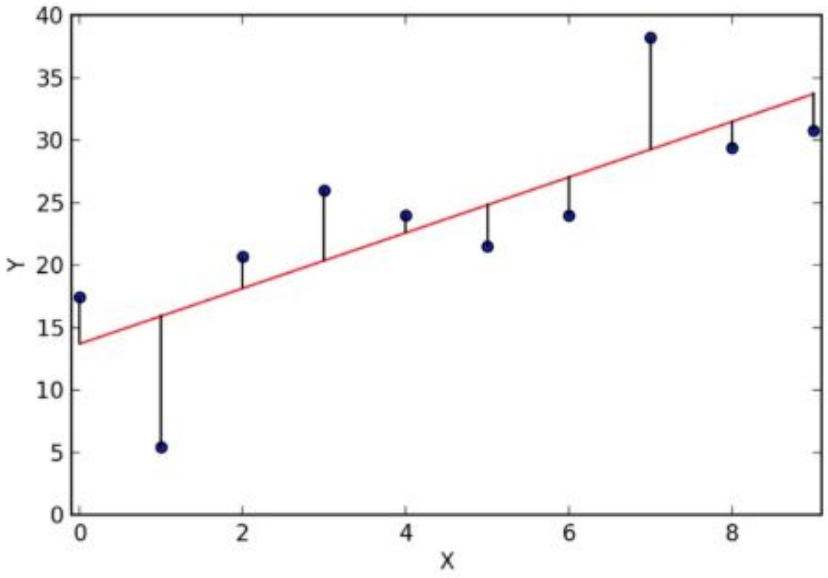
Increase in Employees Turnover

Which variables can be used to understand what is affective the employees' turnover ?

Interactive approach

The Loss

To evaluate the effectiveness of a linear regression model, we must first understand loss. The loss function measures the difference between the predicted and actual results.



Most popular Loss Techniques:

- Mean Square Error
- Mean Absolute Error
- Huber Loss
- Mean Squared Logarithmic Error
- Mean Bias Error

Interactive approach (different variables)

Careful

Correlation is NOT Causation

It's easy to say that there is a correlation between salary and employees' turnover. The regression shows that they are indeed related. But it's an entirely different thing to say that salary *caused* the turnover.

Advice 1
When you see a correlation from a regression analysis, you can't make assumptions. Instead, "You have to go out and see what's happening in the real world. What's the physical mechanism that's causing the relationship?"

Advice 2
START FROM THE BUSINESS SIDE don't tell your data analyst to go out and figure out what is affecting turnover. It's the managers job to identify the factors that you suspect are having an impact and ask your analyst to look at those. Otherwise, you're likely to find relationships that don't really exist.

Advice 3
Focus on what you can change!

Unilever

Video: https://www.hirevue.com/resources/video/unilevers-recruiting-process

Pros

- Dropout rates of over 50% with this traditional approach. Candidates are generally keener to complete a short series of games rather than lengthy tests.
- Instant performance feedback
- Bias Reduction. Furthermore, games elicit more authentic behavior
- games allow for more data points to be collected than traditional multiple choice assessments. Combining this with machine learning techniques will increase predictive validity and, therefore, a more accurate picture of how each jobseeker will perform in their job.
- Time saving

Cons

- Technology might contains bias (needs to be trained)
- A university research study by Greg Sears and Haiyan Zhang found that interviewers tend to form more negative impressions of candidates when interviewed via video. They suggest video interviews are best used as a supplementary screening tool rather than a replacement.
- Employees need to be trained to properly use these technologies

Link to the Case Study

Axtria Case Study

Over **80%** of enterprise data today is **unstructured** , including text, image, and voice and the volume continues to expand.

BUSINESS OBJECTIVE

To provide a comprehensive view of patient reports and deliver critical insights to both healthcare providers, and the company, by:

- Aiding **physician decision making** and **improving quality of care**
- **Identifying a new cancer patient population,** better understand the disease area and treatment patterns
- **Improving patient adherence** through **tracking** patients and their disease progression

AXTRIA'S METHODOLOGY

Big Data Powered by Machine Learning

The focus was on the process of using **cloud and distributed computing** to **upgrade data environments,** experimenting with different tools to **convert images to text,** and applying **Natural Language Processing (NLP)** and **Machine Learning (ML)** to accurately extract information from noisy text.

To address the objectives, a four-step approach was taken:

INGESTION → IMAGE-TO-TEXT TRANSFORMATION → TEXT EXTRACTION → POST EXTRACTION PROCESSING

> *To build machine learning models, we first compiled a training dataset of ~18,000 variants from online databases including the National Health Institute. (Unfortunately, these databases were not comprehensive, so they could not be used as a dictionary for variant detection). Next, features were created and adjusted based on characteristics of each token, which are semantic units in a string divided by blanks and punctuations. Then, we tested multiple algorithms with parameter tuning and feature engineering. Iterations, adjustments and cross-validation show that Gaussian Naïve Bayes algorithm was the most effective at identifying variants.*

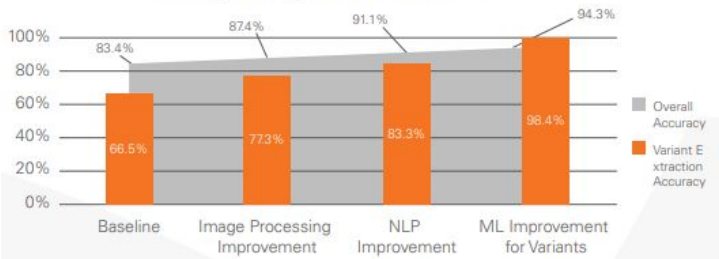
PROJECT CHALLENGE

The work required **accurate extraction and analysis** of information from **~1,000,000 EHR/EMR files** in the form of PDF, image, and XML files, totaling **~7 terabytes of data.**

The key challenges of this process were three-fold:

- The data infrastructure was not set up for storage, rapid access and analysis of unstructured data
- The inherent complexity and variety of the files impeded the analytics process
- Lastly, the quality of the scanned records was low, leading to low-quality image-to-text conversions

4 Stages of Improving Field Extraction



The gray area shows accuracy of all fields, with each stage showing accuracy gain of each improvement. The bar chart indicates the improvements of variant extraction accuracy at each stage.

94% accuracy across different fields, an 11% improvement

32% increase in accuracy, reducing variant error rate to only 1.6%

[Link to the Case Study]

American Express

AI-Driven Fraud Detection

To improve the customer experience and reduce fraud risk, American Express has developed a "fraud model [that] is one of the most advanced in the industry: GenX

- the model evaluates some 8 billion transactions every year and considers many factors within its algorithms, including whether the customer account has been victimized before.
- The model also segments customers into different categories, e.g., frequent travelers, and considers that assignment when evaluating fraud risk. This segmentation allows AmEx to use a single model for both international and US markets
- The model's algorithms assign weights to hundreds of fraud risk indicators and thousands of decision trees, which are subject to continual refinement based on comparisons between predictions and real-world observations

Automating Customer Service Through AI

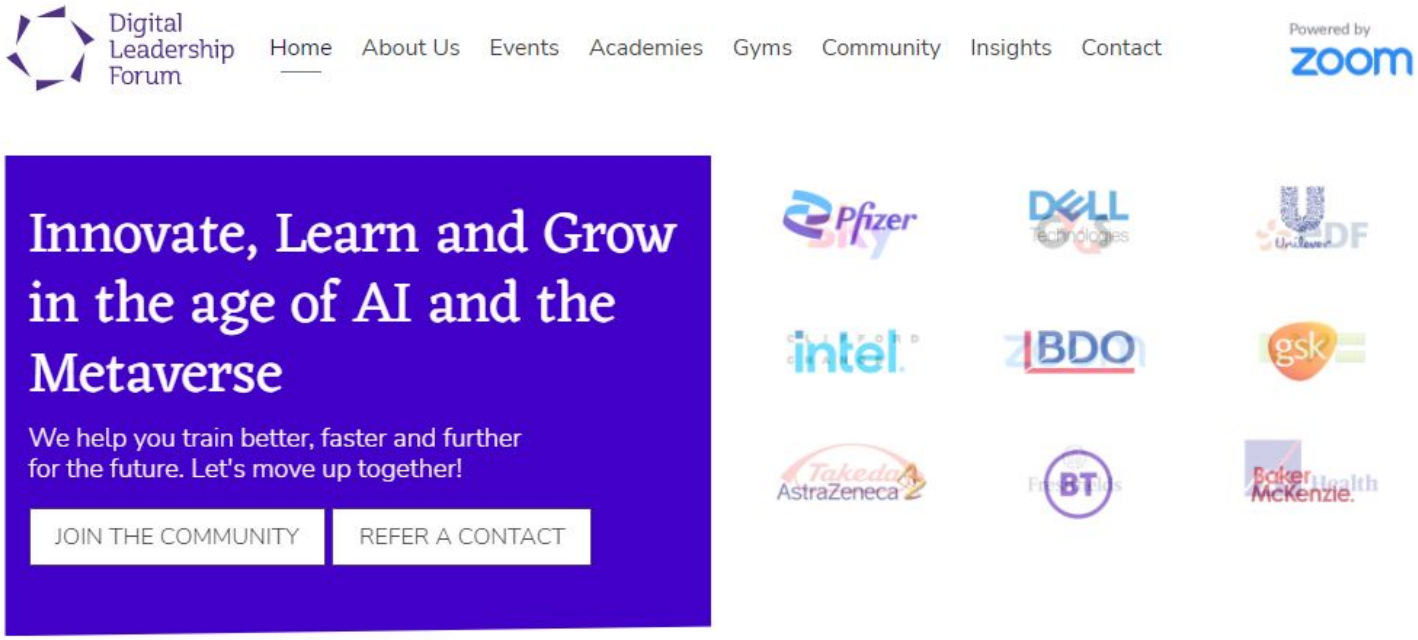
Nine of every ten people assess a company's customer service levels when they decide whether to keep doing business with that company, according to 2020 research published in Microsoft's Global State of Customer Service report.
The credit card issuer has been open in its efforts to integrate machine learning into its customer-facing functions by:

- Transcribing voice to text
- Processing travel bookings
- Automating customer service chat
- Enabling search in the AmEx mobile app
- Classifying emails for delivery to the right departments

For example, in processing travel itineraries, American Express relies on ML to identify the customer's intent—booking travel—and to extract their desired itinerary based on the words they use

Link to the Case Study

Digital Leadership Forum



https://thedigitalleadershipforum.com/

QUESTIONS?