Final Presentation Speaker Notes

Team Introductions - ALL - 1 minute:

- Hayley (sharing screen) - 15 seconds
- Jennifer - 15 seconds
- Maya - 15 seconds
- Lijoy - 15 seconds

Introduce Topic - Hayley - 3 minutes:

- SLIDE 1
    - Today we'll be talking to you about the relationship between the Billboard Hot 100 chart and the Grammy Awards

- SLIDE 3
    - We'll begin by introducing our central question - "Does inclusion on the Billboard Hot 100 chart increase the likelihood that a song will win a Grammy Award?"

    - We chose this topic because the 2021 Grammy Awards had just taken place, and I had prior experience in the music industry which led me to be curious about consumer opinions vs. awards/critical recognition. The rest of my team agreed that this would be a topic with many avenues for exploration, so we began our search for a dataset. - 30 secs

- SLIDE 4
    - We found a detailed Dataset on Kaggle.com that provided data on the Billboard Hot 100 list from 1999-2020. The Dataset also included several CSVs containing information on Spotify listens, Grammy Award Wins, and iTunes purchases. We decided to compare the Billboard Hot 100 Data with Grammy Award Wins, because both the Grammy awards and the Billboard Hot 100 chart are determined by elite committees. Therefore, an additional question for analysis is "Are these committees in agreement on which songs deserve commendation?" - 1 min

    - Once we had chosen our data and central question, the next step was the Data Exploration Phase. We used Excel to examine the data and determine what data cleaning steps would be necessary. Fortunately both the Billboard and Grammy CSVs had consistent formatting and contained a reasonable number of null values. At this point in our exploration we identified additional variables that would be useful in our analysis - Weeks on the Chart, Peak Position, and Genre. The next step was to create visualizations for our initial analysis.

- SLIDE 5
    - Our first chart examines the relationship between the total number of weeks a song was included on the Billboard chart, and the total number of grammy wins. In order to visualize this relationship, we used Excel to create a basic bar chart. We included bins to easily visualize our data. The results of this initial analysis showed that there is little correlation between weeks on the chart, and a grammy win.
    -
- SLIDE 6
    - Our second chart examines the relationship between the peak position that a song reached on the Billboard chart, and the total number of grammy wins. Again we used Excel to create the bar chart, and created the same bins for consistency. Unlike the relationship between weeks on the chart and grammy wins, the results of this initial analysis showed that there did appear to be some correlation between the peak position on the chart, and the likelihood that a song will win a grammy award

    - Now that we had completed our initial data exploration, we were ready to move on to the Data Analysis phase of our project. I will hand it over to Lijoy for an explanation of our Database and Dashboard.

Dashboard - Lijoy - 3 minutes:

- Tools used - Python, Postgres, SQL, Tableau

- How was the Database created?
- We have taken Billboard and Grammy data set for our analysis
- We have used quick database diagram tool to design our database
- As a team we have decided to use Postgres as our database
- Used Python to clean Billboard and Grammy data set
    - Got single instance of each song by dropping duplicates
    - Dropped "weekly_rank", "writing_credits", "lyrics", because they are not relevant for this analysis
    - Totaled number of Grammy Awards for each song to get total number of Grammy Wins
- Loaded these cleaned CSVs into Postgres database using **sqlalchemy**
- Then performed a join between Billboard and Grammy tables to get one complete dataset.
- Later we decided to add genre also into our final dataset to check if genre has any impact on grammy awards, so performed these steps again in an iterative design mode to add genre
- Now our final data set is ready in Postgres - We have used this for ML and in Tableau

for dashboard creation.

-

- Dashboard Walkthrough
    - Explain initial analysis graphs
    - Indicate MLM accuracy score & that Maya would be explaining further
- Interactive features?
    - Walk through at least 2 genres

Machine Learning - Maya - 3 minutes:

- To best serve our needs in predicting whether or not an artist that is on the Billboard Top 100 Chart will win a Grammy, we chose a Logistic Regression model.
- Logistic Regression analysis is appropriate for dealing with binary or yes or no outcomes.
- There are many benefits and a few limitations to using logistic regression models
        - They are simple to implement and make predictions for binary outcomes (yes/no)
        - They are also simple to understand, train, and update with new data to be used in the future
        - However these models cannot be used to solve non-linear problems
        - Such as “How many grammy’s can an artist receive if they are on the billboard top 100 chart?”
        - They require large data set to run analysis
        - And logistic regression models are prone to overfitting
            - Where the model begins to describe the random error in the data rather than the relationships between variables
    - SLIDE 2
    - Feature engineering and Feature Selection were pivotal in maximizing the performance of the machine learning model.
    - We dropped the "weekly_rank", "writing_credits", and "lyrics" variables because they were not informative for our logistic regression analysis.
    - Dropping these unnecessary variables allowed for our machine learning model to easily read through the independent variables and
    - Increased our overall Accuracy Score from 84.62% to 94.12% which I will discuss in more detail in a moment.

    - For the training and testing of our machine learning model
    - We used X to predict y
    - X, or features, was created by dropping the “artists” and “name” columns from the DataFrame.
    - And y or the output is the “GrammyAward” column expressing which artists received Grammy awards,

        - We utilized the train_test_split module to split X and y into four training and testing sets: X_train, X_test, y_train, y_test.

        - We then compare the actual outcome values from the test set against the model's predicted values.

        - For our analysis, the y_test set contained the outcomes of whether or not an artist that is on the Billboard Top 100 Chart won a Grammy from our original dataset

        - The model's predictions, *y_pred,* were compared with these actual values in the, y_test set

- Description of current accuracy score
    - At first glance the predictions from running our machine learning model seemed to be accurate but we generate an accuracy score to determine the percentage of predictions that were correct
    - The accuracy of the machine learning model is 94.12%. This shows that the MLM will accurately predict whether an artist will receive a grammy based on their Billboard Top 100 Chart performance 94.12% of the time.
    - I will hand it over to Jennifer to expand on our Results.

Results - Jennifer - 3 minutes:
Results and Other Considerations

- Final model accuracy of 94.12%.
    - Accept the hypothesis that Grammy winners can be predicted using Billboard Charts.
    - Minimal correlation showed in initial analysis and basic Excel regression.
    - Discus Michael Buble - 4x Grammy, no Top 20, no more than 10 weeks
    - Discuss Gangnam Style - no Grammy, #2 for 4 weeks, total of 35 weeks on chart
- Incomplete dataset due to nature of the Billboard Chart.
    - New song on chart - may not provide actual weeks on chart
    - Song not a peak position within dataset
    - Locate additional information and datasets required
- Changes to Billboard Compilation
    - Originally compiled via committee with data from radio plays and purchases
    - Modified after YouTube success of Gangnam style to include heavier emphasis on digital sales and YouTube views of music videos.
    - Changed in 2017 to permanently include these items and nearly eliminate actual CD sales.

Additions

- Reiteration of accuracy score of 94.12% with datasets

        - Expansion to include datasets from Spotify streaming, Pandora plays, Apple purchases, Amazon music, or many others that are popular in today's culture.
    - Automatic scraper to add weekly Billboard Charts.
        - Not incorporated due to inconsistencies in data format
        - Failed Google and StackOverflow skills
    - All Billboards and Grammy Winners
        - Expand charts to historical information, including Billboard Hot 100 from origination of chart or Grammy Winners from original Grammys
        - Test and train to determine a category winner, not just a Grammy winner.
            - Expansion of categories throughout the Grammy time period (1959 - 14, 2021 - 84).

Questions