Option #2: Capstone Project—Final Report and Slide Presentation: Non-U.S. Organization

Scott Miner

Colorado State University – Global Campus

Abstract

Music streaming services are changing how consumers listen to music. Determining the attributes that lead to trending songs can help services create better user experiences and more effective marketing techniques. This paper analyzes two datasets containing songs' characteristics to determine whether there are any correlations between these characteristics and a song's popularity on Spotify and whether it is possible to predict a song's popularity or chances of becoming a hit on the Billboard Hot 100 Charts. Various statistical tests were used to answer these questions. The results indicate that songs comprised of specific features are significantly more popular on Spotify than others and that some predictor variables are significantly correlated with the criterion variables. Logistic regression using a backward selection technique was shown to outperform all other models in the classification problem, though all predictive models demonstrated substandard prediction capabilities.

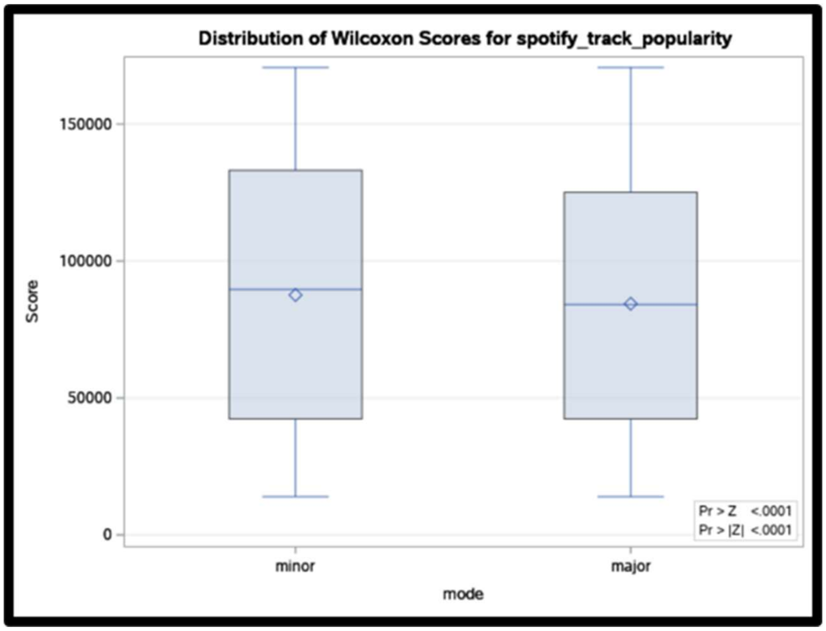


Figure 1. Boxplots showing the popularity score distributions for songs grouped by song mode in the first dataset

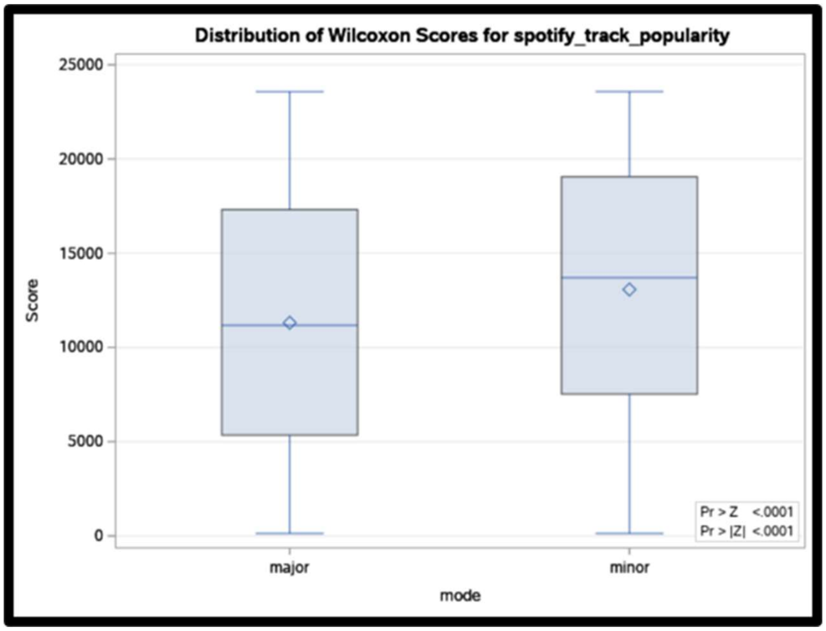


Figure 2. Boxplots showing the popularity score distributions for songs grouped by song mode in the second dataset

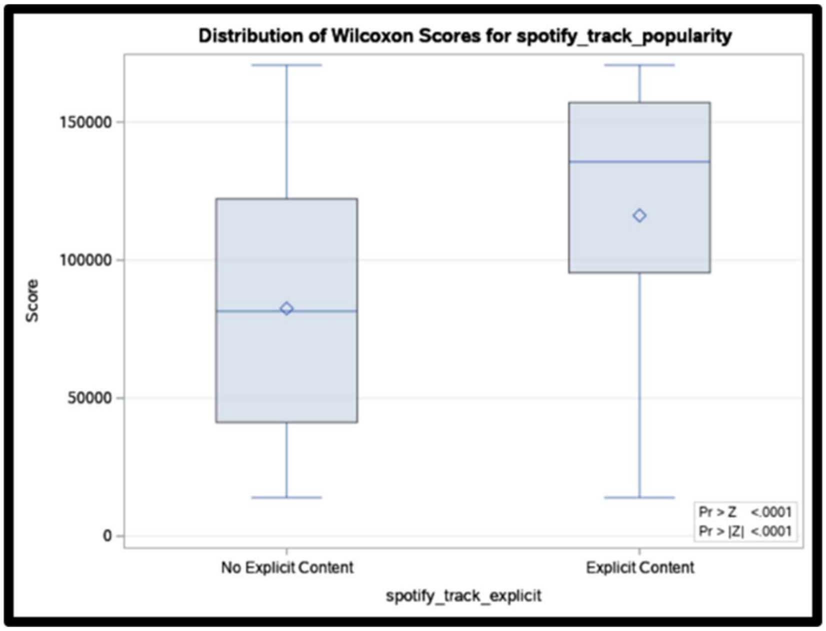


Figure 3. Boxplots showing the popularity score distributions for songs grouped by explicit lyrical content in the first dataset

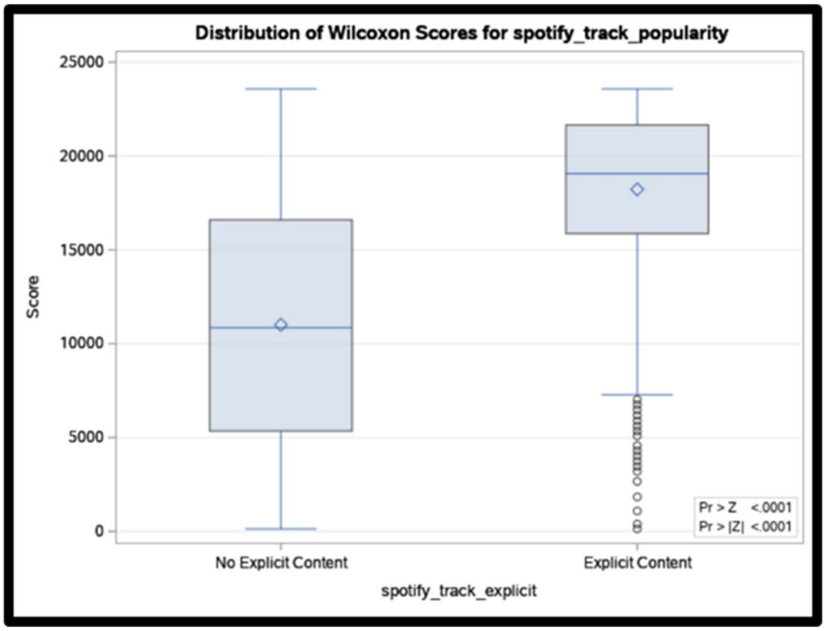


Figure 4. Boxplots showing the popularity score distributions for songs grouped by explicit lyrical content in the second dataset

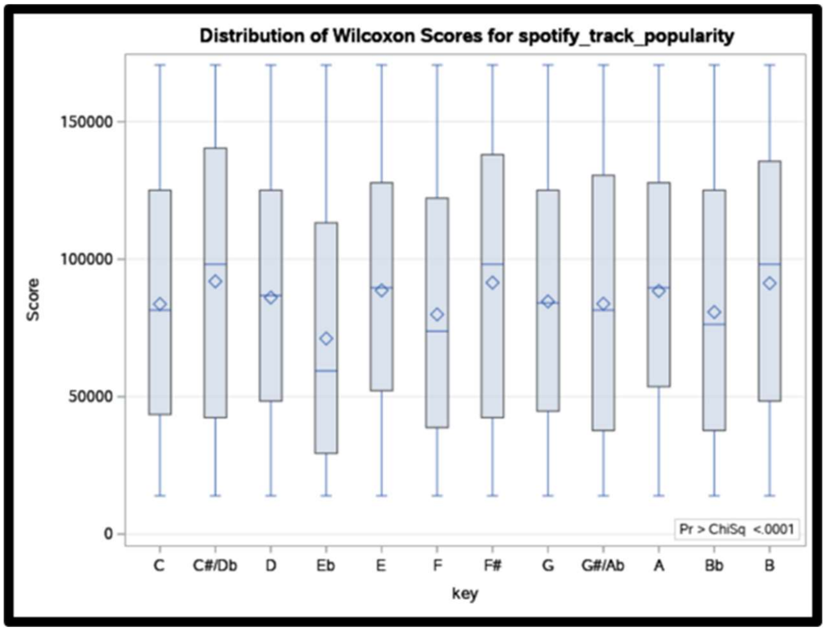


Figure 5. Boxplots showing the popularity score distributions for songs grouped by song key in the first dataset

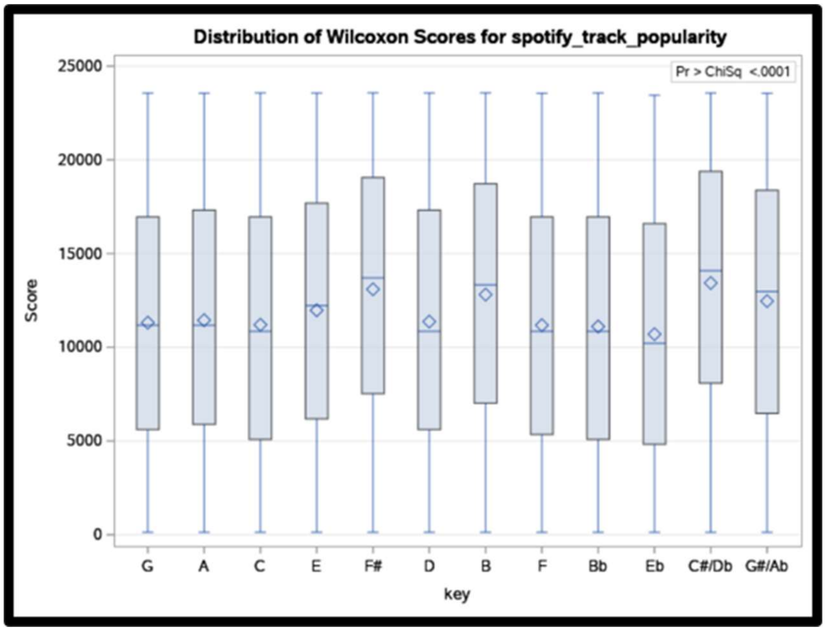
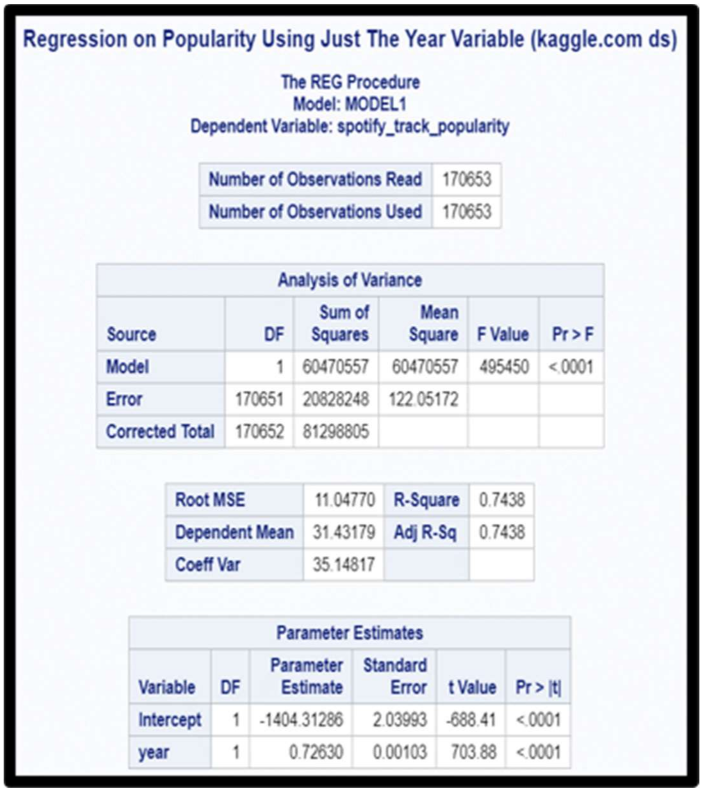


Figure 6. Boxplots showing the popularity score distributions for songs grouped by song key in the second dataset



Regression on Popularity Using Just The Year Variable (kaggle.com ds)

The REG Procedure
Model: MODEL1
Dependent Variable: spotify_track_popularity

Number of Observations Read	170653
Number of Observations Used	170653

Analysis of Variance

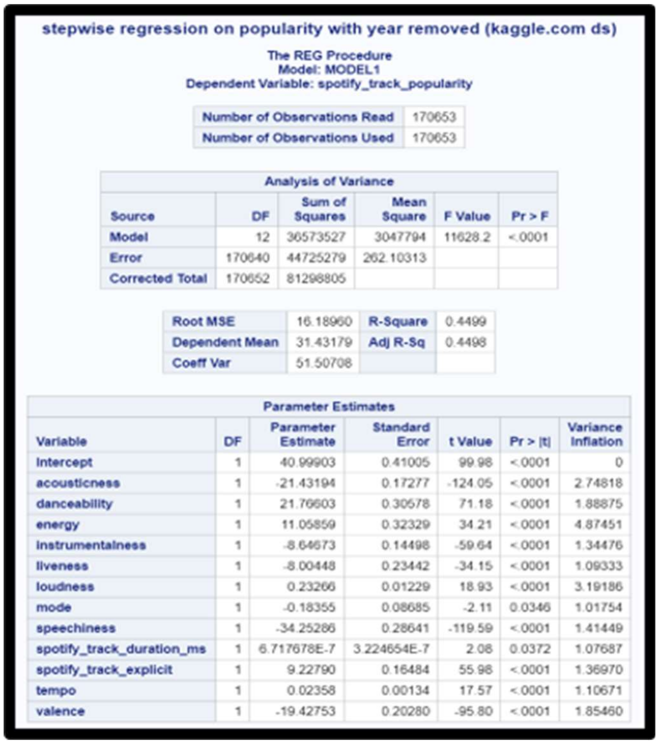
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	60470557	60470557	495450	<.0001
Error	170651	20828248	122.05172		
Corrected Total	170652	81298805			

Root MSE	11.04770	R-Square	0.7438
Dependent Mean	31.43179	Adj R-Sq	0.7438
Coeff Var	35.14817		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1404.31286	2.03993	-688.41	<.0001
year	1	0.72630	0.00103	703.88	<.0001

Figure 7. Results of the simple linear regression model



stepwise regression on popularity with year removed (kaggle.com ds)

The REG Procedure
Model: MODEL1
Dependent Variable: spotify_track_popularity

Number of Observations Read	170653
Number of Observations Used	170653

Analysis of Variance

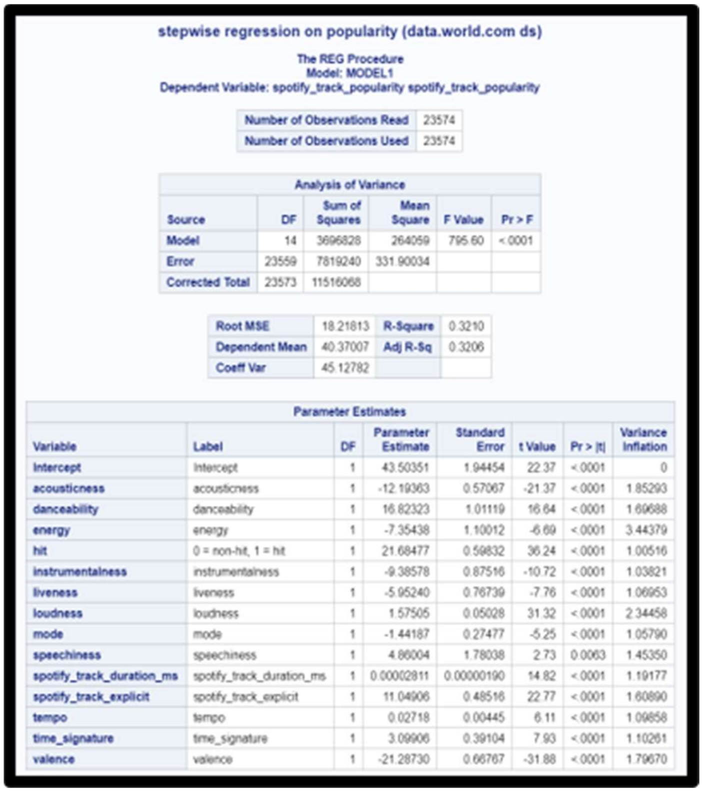
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	36573527	3047794	11628.2	<.0001
Error	170640	44725279	262.10313		
Corrected Total	170652	81298805			

Root MSE	16.18960	R-Square	0.4499
Dependent Mean	31.43179	Adj R-Sq	0.4498
Coeff Var	51.50708		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	40.99903	0.41005	99.98	<.0001	0
acousticness	1	-21.43194	0.17277	-124.05	<.0001	2.74818
danceability	1	21.76603	0.30578	71.18	<.0001	1.88875
energy	1	11.05859	0.32329	34.21	<.0001	4.87451
instrumentalness	1	-8.64673	0.14498	-59.64	<.0001	1.34476
liveness	1	-8.00448	0.23442	-34.15	<.0001	1.09333
loudness	1	0.23266	0.01229	18.93	<.0001	3.19186
mode	1	-0.18355	0.08685	-2.11	0.0346	1.01754
speechiness	1	-34.25286	0.28641	-119.59	<.0001	1.41449
spotify_track_duration_ms	1	6.717678E-7	3.224654E-7	2.08	0.0372	1.07687
spotify_track_explicit	1	9.22790	0.16484	55.98	<.0001	1.36970
tempo	1	0.02358	0.00134	17.57	<.0001	1.10671
valence	1	-19.42753	0.20280	-95.80	<.0001	1.85460

Figure 8. Results of the multiple linear regression model on the first dataset when the year is removed



stepwise regression on popularity (data.world.com ds)

The REG Procedure
Model: MODEL1
Dependent Variable: spotify_track_popularity spotify_track_popularity

Number of Observations Read	23574
Number of Observations Used	23574

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	3696828	264059	795.60	<.0001
Error	23559	7819240	331.90034		
Corrected Total	23573	11516068			

Root MSE	18.21813	R-Square	0.3210
Dependent Mean	40.37007	Adj R-Sq	0.3206
Coeff Var	45.12782		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	43.50351	1.94454	22.37	<.0001	0
acousticness	acousticness	1	-12.19363	0.57067	-21.37	<.0001	1.85293
danceability	danceability	1	16.82323	1.01119	16.64	<.0001	1.69688
energy	energy	1	-7.35438	1.10012	-6.69	<.0001	3.44379
hit	0 = non-hit, 1 = hit	1	21.68477	0.59832	36.24	<.0001	1.00516
instrumentalness	instrumentalness	1	-9.38578	0.87516	-10.72	<.0001	1.03821
liveness	liveness	1	-5.95240	0.76739	-7.76	<.0001	1.06953
loudness	loudness	1	1.57505	0.05028	31.32	<.0001	2.34458
mode	mode	1	-1.44187	0.27477	-5.25	<.0001	1.05790
speechiness	speechiness	1	4.86004	1.78038	2.73	0.0063	1.45350
spotify_track_duration_ms	spotify_track_duration_ms	1	0.00002811	0.00000190	14.82	<.0001	1.19177
spotify_track_explicit	spotify_track_explicit	1	11.04906	0.48516	22.77	<.0001	1.60890
tempo	tempo	1	0.02718	0.00445	6.11	<.0001	1.09858
time_signature	time_signature	1	3.09906	0.39104	7.93	<.0001	1.10261
valence	valence	1	-21.28730	0.66767	-31.88	<.0001	1.79670

Figure 9. Results of the multiple linear regression model on the second dataset

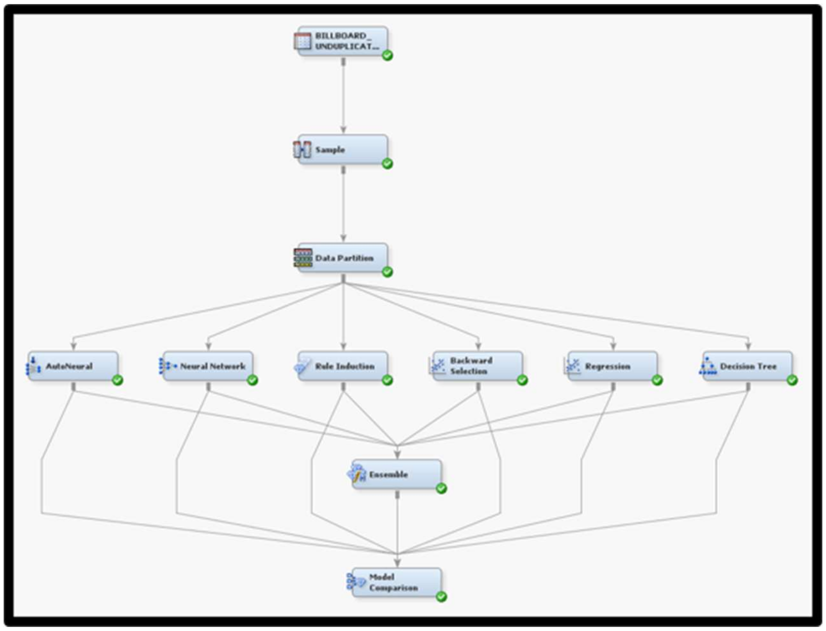
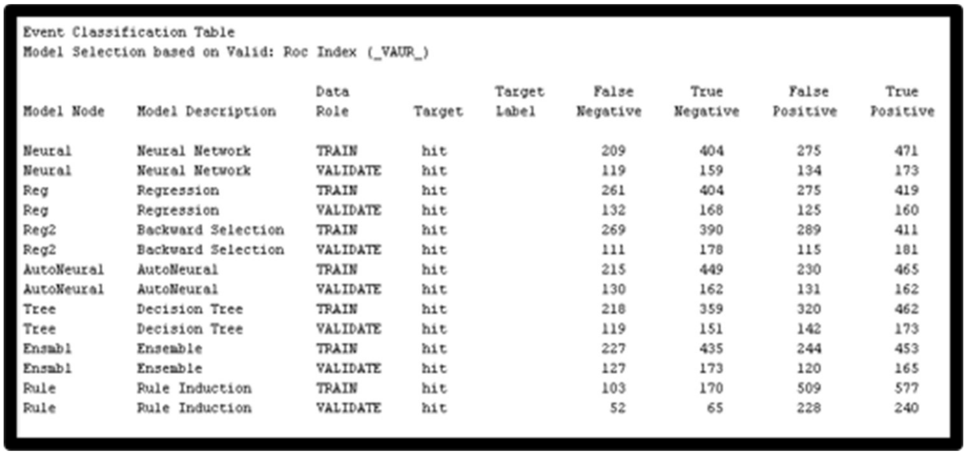
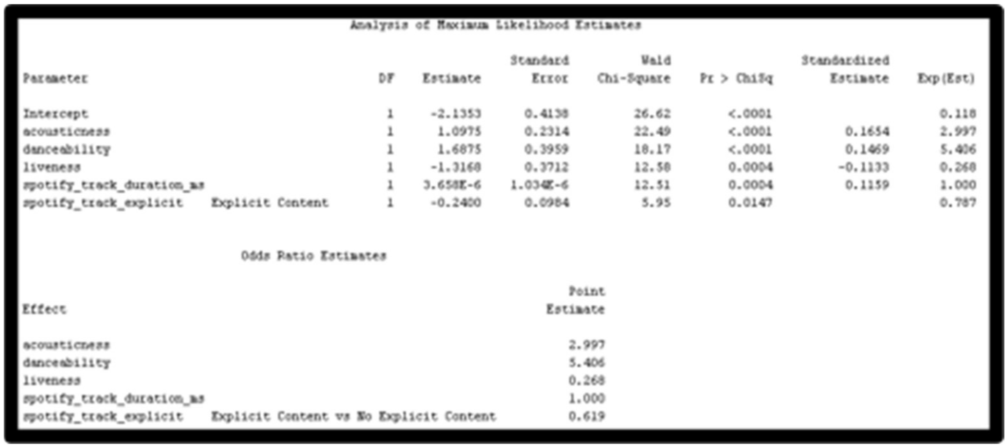


Figure 10. Workflow diagram in SAS Enterprise Miner



```
Event Classification Table
Model Selection based on Valid: Roc Index (_VAUR_)

                                 Data                 Target    False     True     False     True
Model Node   Model Description   Role        Target   Label    Negative  Negative  Positive  Positive

Neural       Neural Network      TRAIN        hit                 209       404       275       471
Neural       Neural Network      VALIDATE     hit                 119       159       134       173
Reg          Regression          TRAIN        hit                 261       404       275       419
Reg          Regression          VALIDATE     hit                 132       168       125       160
Reg2         Backward Selection  TRAIN        hit                 269       390       289       411
Reg2         Backward Selection  VALIDATE     hit                 111       178       115       181
AutoNeural   AutoNeural          TRAIN        hit                 215       449       230       465
AutoNeural   AutoNeural          VALIDATE     hit                 130       162       131       162
Tree         Decision Tree       TRAIN        hit                 218       359       320       462
Tree         Decision Tree       VALIDATE     hit                 119       151       142       173
Ensmbl       Ensemble            TRAIN        hit                 227       435       244       453
Ensmbl       Ensemble            VALIDATE     hit                 127       173       120       165
Rule         Rule Induction      TRAIN        hit                 103       170       509       577
Rule         Rule Induction      VALIDATE     hit                  52        65       228       240
```

Figure 11. Results of the model comparison node



```
                              Analysis of Maximum Likelihood Estimates

                                                  Standard          Wald                   Standardized
Parameter                          DF   Estimate      Error    Chi-Square    Pr > ChiSq        Estimate    Exp(Est)

Intercept                           1    -2.1353     0.4138         26.62        <.0001                       0.118
acousticness                        1     1.0975     0.2314         22.49        <.0001          0.1654       2.997
danceability                        1     1.6875     0.3959         18.17        <.0001          0.1469       5.406
liveness                            1    -1.3168     0.3712         12.58        0.0004         -0.1133       0.268
spotify_track_duration_ms           1   3.658E-6   1.034E-6         12.51        0.0004          0.1159       1.000
spotify_track_explicit    Explicit Content    1    -0.2400     0.0984          5.95        0.0147                       0.787


                         Odds Ratio Estimates

                                                               Point
Effect                                                      Estimate

acousticness                                                   2.997
danceability                                                   5.406
liveness                                                       0.268
spotify_track_duration_ms                                      1.000
spotify_track_explicit    Explicit Content vs No Explicit Content    0.619
```

Figure 12. Results from the logistic regression with backward selection technique

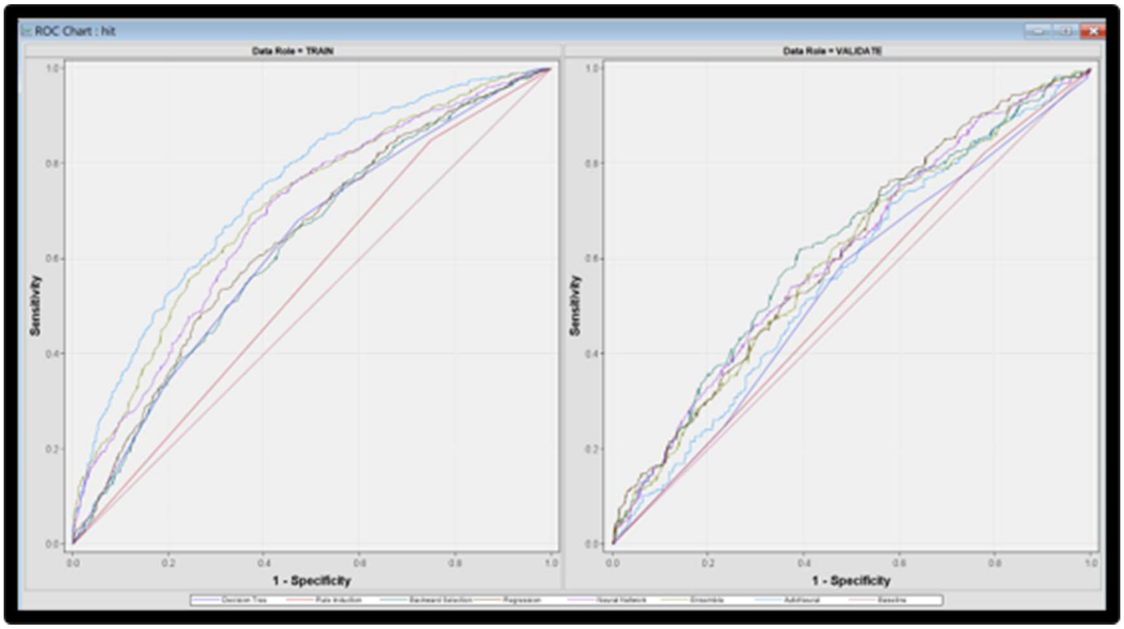
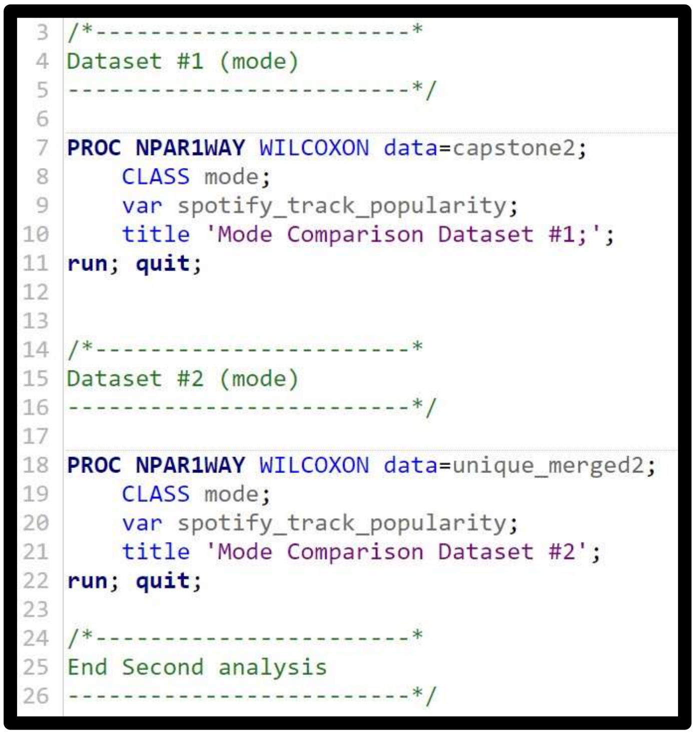
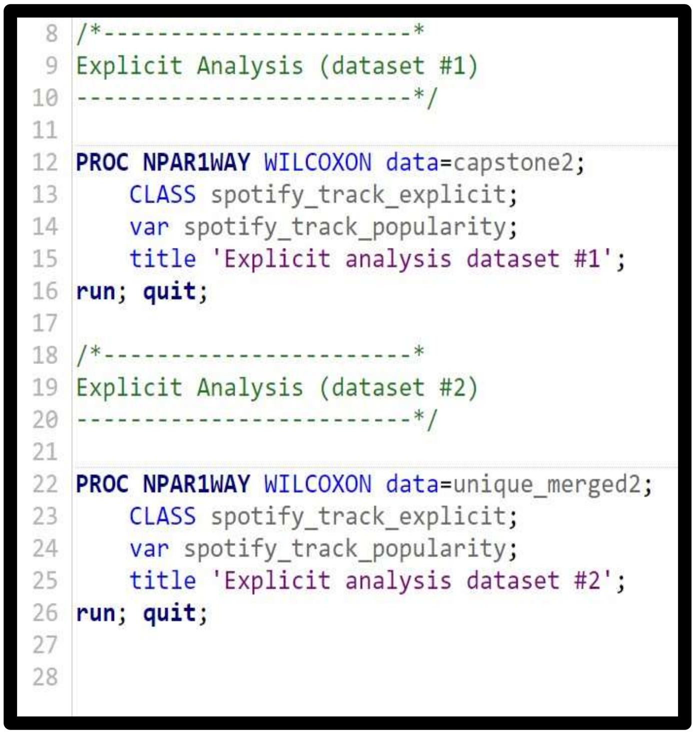


Figure 13. ROC curve for training and validation datasets



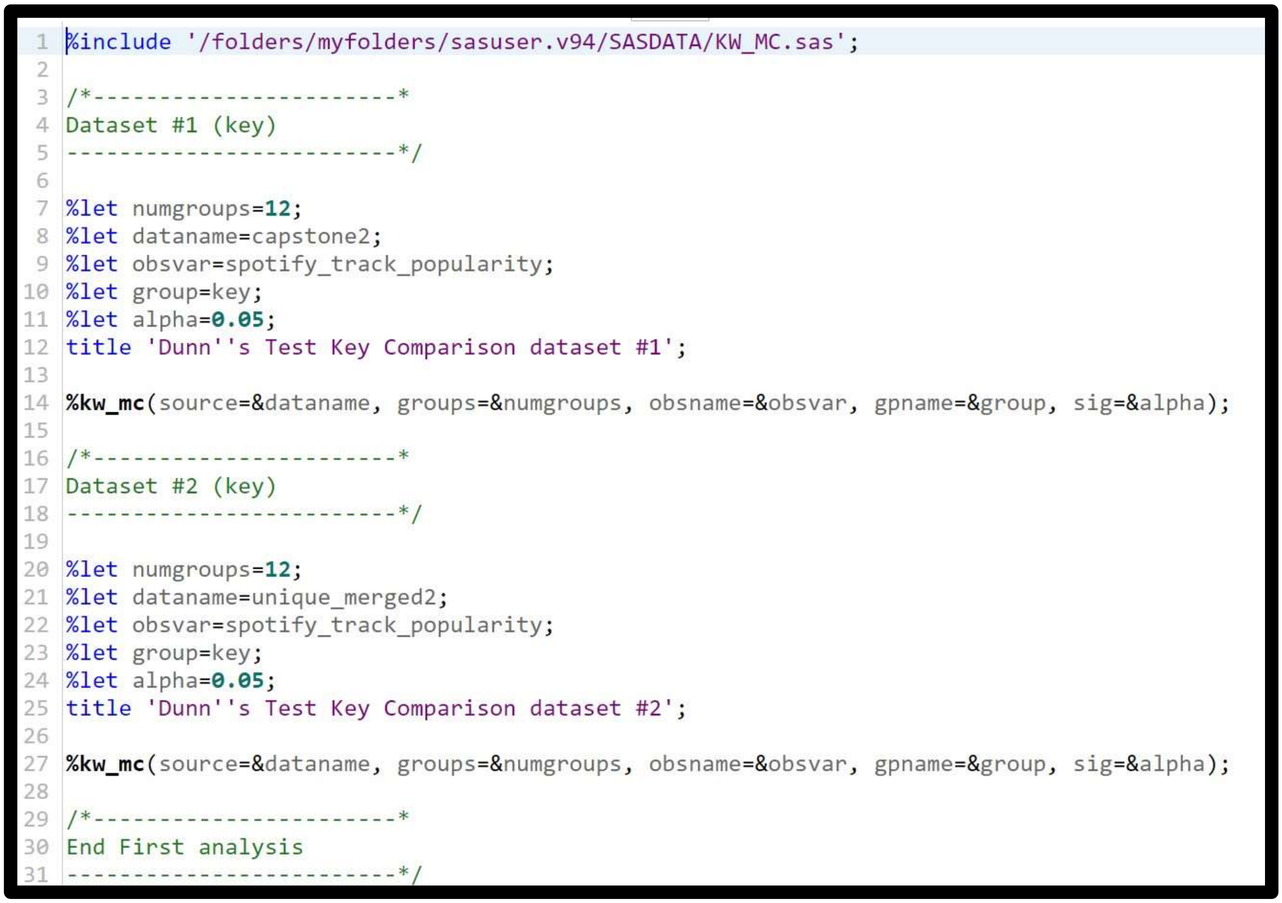
```
/*-----------------------*
Dataset #1 (mode)
-------------------------*/

PROC NPAR1WAY WILCOXON data=capstone2;
    CLASS mode;
    var spotify_track_popularity;
    title 'Mode Comparison Dataset #1;';
run; quit;


/*-----------------------*
Dataset #2 (mode)
-------------------------*/

PROC NPAR1WAY WILCOXON data=unique_merged2;
    CLASS mode;
    var spotify_track_popularity;
    title 'Mode Comparison Dataset #2';
run; quit;

/*-----------------------*
End Second analysis
-------------------------*/
```

Figure 14. SAS Code comparing song popularity distributions by song mode



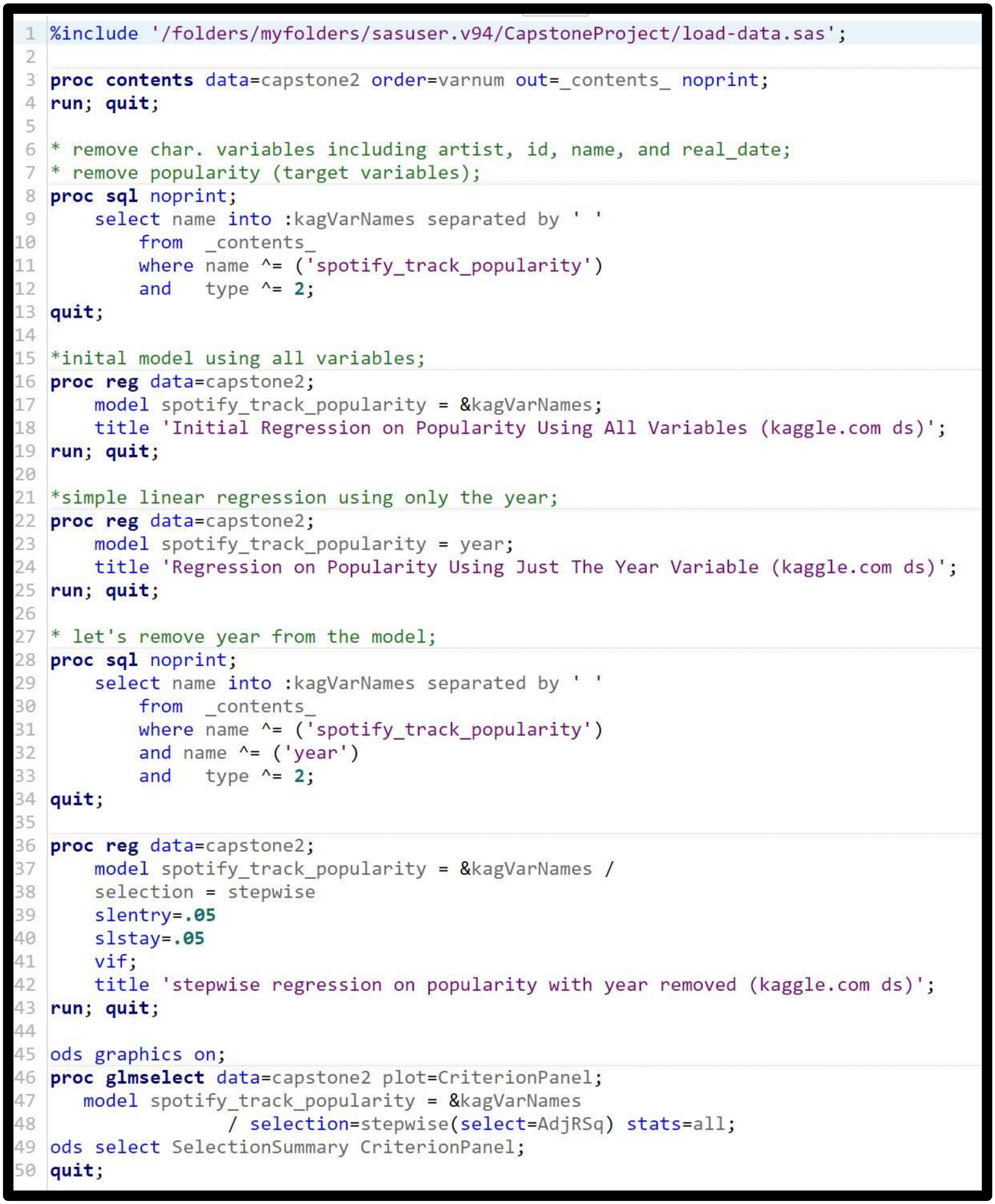
```
/*-----------------------*
Explicit Analysis (dataset #1)
-------------------------*/

PROC NPAR1WAY WILCOXON data=capstone2;
    CLASS spotify_track_explicit;
    var spotify_track_popularity;
    title 'Explicit analysis dataset #1';
run; quit;

/*-----------------------*
Explicit Analysis (dataset #2)
-------------------------*/

PROC NPAR1WAY WILCOXON data=unique_merged2;
    CLASS spotify_track_explicit;
    var spotify_track_popularity;
    title 'Explicit analysis dataset #2';
run; quit;
```

Figure 15. SAS code comparing song popularity distributions by explicit lyrical content



```
%include '/folders/myfolders/sasuser.v94/SASDATA/KW_MC.sas';

/*-----------------------*
Dataset #1 (key)
-------------------------*/

%let numgroups=12;
%let dataname=capstone2;
%let obsvar=spotify_track_popularity;
%let group=key;
%let alpha=0.05;
title 'Dunn''s Test Key Comparison dataset #1';

%kw_mc(source=&dataname, groups=&numgroups, obsname=&obsvar, gpname=&group, sig=&alpha);

/*-----------------------*
Dataset #2 (key)
-------------------------*/

%let numgroups=12;
%let dataname=unique_merged2;
%let obsvar=spotify_track_popularity;
%let group=key;
%let alpha=0.05;
title 'Dunn''s Test Key Comparison dataset #2';

%kw_mc(source=&dataname, groups=&numgroups, obsname=&obsvar, gpname=&group, sig=&alpha);

/*-----------------------*
End First analysis
-------------------------*/
```

Figure 16. SAS code comparing song popularity distributions by song key



```
%include '/folders/myfolders/sasuser.v94/CapstoneProject/load-data.sas';

proc contents data=capstone2 order=varnum out=_contents_ noprint;
run; quit;

* remove char. variables including artist, id, name, and real_date;
* remove popularity (target variables);
proc sql noprint;
    select name into :kagVarNames separated by ' '
        from  _contents_
        where name ^= ('spotify_track_popularity')
        and   type ^= 2;
quit;

*inital model using all variables;
proc reg data=capstone2;
    model spotify_track_popularity = &kagVarNames;
    title 'Initial Regression on Popularity Using All Variables (kaggle.com ds)';
run; quit;

*simple linear regression using only the year;
proc reg data=capstone2;
    model spotify_track_popularity = year;
    title 'Regression on Popularity Using Just The Year Variable (kaggle.com ds)';
run; quit;

* let's remove year from the model;
proc sql noprint;
    select name into :kagVarNames separated by ' '
        from  _contents_
        where name ^= ('spotify_track_popularity')
        and name ^= ('year')
        and   type ^= 2;
quit;

proc reg data=capstone2;
    model spotify_track_popularity = &kagVarNames /
    selection = stepwise
    slentry=.05
    slstay=.05
    vif;
    title 'stepwise regression on popularity with year removed (kaggle.com ds)';
run; quit;

ods graphics on;
proc glmselect data=capstone2 plot=CriterionPanel;
   model spotify_track_popularity = &kagVarNames
                / selection=stepwise(select=AdjRSq) stats=all;
ods select SelectionSummary CriterionPanel;
quit;
```

Figure 17. SAS code performing simple and multiple linear regression on the popularity variable using all other dataset variables

Table 1. Data dictionary for the Spotify dataset

Variable	Type	Definition
Acousticness	Float	A confidence measure from 0.0 to 1.0, indicating whether a track is acoustic.
Danceability	Float	A track's suitability for dancing from 0.0 to 1.0, based on a combination of factors.
Energy	Float	The perceptual measure of a track's intensity and activity on a scale of 0.0 to 1.0.
Hit	Bool	A binary variable indicating whether a track achieved a top spot on the Billboard Hot 100 Chart.
Instance	Int	A value between 1 and 9 indicating how many times a song has appeared on the Hot 100 singles chart.
Instrumentalness	Float	A measure from 0.0 to 1.0 of a track's vocal content.
Key	Int	Estimates a track's over all key using pitch class notation (e.g., 0 = C, 1 = C#/Db, 2 = D).
Liveness	Float	Describes the likelihood that a track was recorded with a live audience on a scale of 0.0 to 1.0.
Loudness	Float	A measurement from -60 to 0 in decibels (dB) of the magnitude of the auditory sensation a track produces.
Mode	Bool	The scale from which a track derives its melodic content. Values of 1 represent the major scale, whereas 0 represents the minor scale.
Peak_Position	Int	A value between 1 and 100 indicating the song's peak position on the Billboard charts as of the corresponding week.
Performer	Char	The name of the artist who released the track.

Previous_Week_Position	Int	A number between 1 and 100 indicating the song's previous position on the Hot 100 singles chart.
Release_Date	YYYY-MM-DD	The date that a track was released.
Song	Char	The name of the track.
SongId	Char	A concatenation of the song name and performer name.
Speechiness	Float	A value from 0.0 to 1.0 representing the presence of spoken words in a track. Values above 0.66 indicate tracks most likely composed entirely of spoken words.
Spotify_Track_Album	Char	The album from which the track was obtained.
Spotify_Track_Duration_Ms	Int	The track's duration in milliseconds.
Spotify_Track_Explicit	Bool	Indicates the presence (1) or absence (0) of explicit language.
Spotify_Track_Genre	Char	The genre of the track.
Spotify_Track_Id	Char	The track's Spotify ID.
Spotify_Track_Popularity	Float	A value between 0 and 100 based on total track plays and their recency.
Spotify_Track_Preview_Url	Char	A URL to a 30-second preview of the song.
Tempo	Float	A measure of the track's speed in beats per minute (BPM).
Time_Signature	Int	A value ranging from 0 to 5 that indicates the time signature of the track.
Url	Char	Billboard chart URL.
Valence	Float	The musical positiveness that a track conveys on a scale of 0.0 to 1.0.
Week_Position	Int	An integer between 1 and 100 reflecting the song's position on the Hot 100 singles chart for the corresponding week.
WeekId	MM/DD/YYYY	The week the song was on the Hot 100 singles chart.
Weeks_On_Chart	Int	Specifies the number of weeks on the chart as of the corresponding week.

Year	YYYY	The year a song was released.

Table 2. Model comparison metrics

Model	Accuracy	MR	Sensitivity	Specificity	Precision
Neural	0.57	0.43	0.59	0.54	0.56
Regression	0.56	0.44	0.55	0.57	0.56
Backward Selection	0.61	0.39	0.62	0.61	0.61
AutoNeural	0.55	0.45	0.55	0.55	0.55
Decision Tree	0.55	0.45	0.59	0.52	0.55
Ensemble	0.58	0.42	0.57	0.59	0.58
Rule Induction	0.52	0.48	0.82	0.22	0.51

Table 3. Additional model comparison metrics

Model	Geometric Mean	Discrimant Power	Balanced Accuracy	MCC	Youden's Index	PLR	NLR
Neural	**0.57**	**0.13**	**0.16**	**0.14**	**0.14**	**1.30**	**0.75**
Regression	**0.56**	**0.12**	**0.16**	**0.12**	**0.12**	**1.28**	**0.79**
Backward Regression	**0.61**	**0.22**	**0.19**	**0.23**	**0.23**	**1.58**	**0.63**

AutoNeural	**0.55**	**0.10**	**0.15**	**0.11**	**0.11**	**1.24**	**0.81**
Decision Tree	**0.55**	**0.10**	**0.15**	**0.11**	**0.11**	**1.22**	**0.79**
Ensemble	**0.58**	**0.15**	**0.17**	**0.16**	**0.16**	**1.38**	**0.74**
Rule Induction	**0.43**	**0.07**	**0.09**	**0.05**	**0.04**	**1.06**	**0.80**

Contents

Option #2: Capstone Project Final Report: Non-U.S. Organization

For his *Portfolio Project* in *MIS581-Capstone Business Intelligence and Data Analytics*, the student submits his final report for his Capstone Project, including all project components from previous Critical Thinking assignments. The student describes the results of his analysis and explains whether he was able to prove or disprove his hypotheses.

Introduction

The author chose to analyze a dataset that would benefit *Spotify* for his Capstone Project. Founded in 2006, Spotify is an international 'on-demand' music-streaming service headquartered in Stockholm, Sweden (Fleischer & Snickars, 2017). The difference between on-demand streaming services and streaming radio services like Pandora is that on-demand streaming services are interactive instead of non-interactive, allowing users to choose what they want to hear (Marshall, 2015). Spotify initially positioned itself as a legal alternative to piracy-based applications like the PirateBay by signing licensing agreements with major record label companies. The service was initially available in a handful of European countries before later launching in the U.S. in July 2011 (Wikhamn & Knights, 2016).

Spotify offers desktop, web, and mobile applications that allow users to search for, play, and download music. Users can also create playlists to earmark tracks for future retrieval and share with others (Haupt & Shelley, 2012). The company offers two membership tiers: (a) a free version, which limits playback functionality and incorporates advertisements into the listening experience; and (b) a premium version, which costs under $10 a month and provides an ad-free listening experience, as well as additional features for subscribers (Eltringham, 2017). From 2011 to 2019, the number of employees at Spotify grew from 311 to 4,405 (*Spotify Employees Count 2018*, n.d.). In 2019, Spotify reported total annual revenue of $7.44 billion (Fielding, 2020). As of September 30, 2020, the organization had 320 million monthly active users, 144 million subscribers, 60+ million tracks, and 1.9+ million podcasts in 92 countries (*Spotify — Company Info*, 2020).

Research Hypothesis

A hypothesis is an educated guess about the relationship between two or more variables, articulated in the form of a testable statement (O'Leary, 2017). The null hypothesis represents what one would expect to see if there were no correlation between two or more variables, whereas the alternative hypothesis represents a researcher's hunch about the variables in question. One question this research seeks to answer is, "Are there any differences among the distributions of popularity scores for songs grouped by song mode (e.g., major or minor), explicit lyrical content, or song key?" The null hypothesis, H_0, states the distributions of popularity scores are the same for these groups. On the other hand, the alternate hypothesis, H_a, states that the distributions of popularity scores are not all the same.

Secondly, the research seeks to answer, "How successfully can one predict a song's popularity score on Spotify and how successfully can one predict if a song will achieve a number one spot on the Billboard Hot 100 Chart, given any other characteristics in the dataset?" The null hypothesis, H_0, states there is no statistically significant relationship between a song's popularity score on the Spotify platform or a song's chances of achieving a number one spot on the Billboard Hot 100 Chart and any other song characteristics in the dataset. The alternate hypothesis, H_a, states there is a statistically significant relationship between these variables.

Objectives

This study aims to identify any differences among the distributions of Spotify popularity scores for songs grouped by song mode, explicit lyrical content, and song key. Additionally, the study aims to identify how successfully one can predict a song's popularity score on Spotify and a song's chances of becoming a hit on the Billboard Hot 100 Chart given a set of song features.

Overview of Study

The remainder of this paper is divided into the following sections: (a) literature review, (b) research design, (c) findings, (d) conclusion, and (e) recommendations. The literature review presents an overview of past research in Hit Song Science (HSS). The research design section presents the study's methodology, methods, limitations, and ethical concerns. Following this is the findings section, which presents answers to the critical research questions outlined above. The conclusion summarizes the research processes, and the recommendations section provides future research suggestions.

Literature Review

Dhanaraj and Logan (2005) present one of the first investigations into HSS. The researchers built support vector machines (SVMs) and boosting classifiers using acoustic and lyric-based song features to distinguish top 1 hits from non-hits. The authors found their classifiers to be better than random. Pachet and Roy (2008) used SVMs to determine the validity of the former researchers' claims. Pachet and Roy found they could not develop an accurate classification system for *low, medium,* and *high* popularity. The authors attributed Dhanaraj's and Logan's findings to spurious data. Ni *et al.* (2011) achieved more optimistic results, finding they could predict whether a song would reach a top 5 position on the U.K. top 40 singles chart, attributing their success to the novel audio features used to build the study's shifting perceptron model. Finally, Herremans *et al.* (2014) built and compared several models capable of predicting top 10 dance hits using features like those obtained by Ni *et al.* The authors found logistic regression performed best for this classification problem.

Research Design

Fitzgerald *et al.* (2004) write that researchers often employ correlational designs to explore relationships between variables they cannot manipulate. Researchers conduct correlational designs for explanatory and predictive purposes. In predictive correlational designs, researchers gather information on how one or more predictor variables correlate with a criterion variable. This study employed a *predictive correlational design* to determine how a song's characteristics correspond with its popularity score on Spotify and its potential to achieve a number one spot on the Billboard Hot 100 Charts.

Methodology

O'Leary (2017) describes two research methodologies, quantitative and qualitative. The quantitative approach relies on numerical data, utilizes the scientific method, and tests hypotheses, whereas the qualitative approach relies on data not easily quantified. This study analyzed quantitative data using statistical tests, employing a quantitative methodology.

Methods

The author chose to analyze a dataset from Spotify because he has been subscribing to the service since 2017. Additionally, the author found several datasets available online pulled from the Spotify Web API. The first dataset is available on Kaggle.com, a platform for machine learning competitions (Narayanan *et al.*, 2011). The dataset contains a host of musical attributes attributed to more than 170,000 songs in the Spotify catalog released between 1921 and 2020 (*Spotify Dataset 1921-2020, 160k+ Tracks*, n.d.). The researcher also obtained a second dataset available on data.world containing musical attributes for every song on the Billboard Hot 100 Chart released between August 2, 1958, and December 26, 2020 (Miller, n.d.).

Each row in each dataset corresponds to a single track. Table 1 presents a data dictionary for all variables in both datasets, including each variable's type and definition. The researcher primarily used SAS and SAS Enterprise Miner to analyze these datasets. The datasets were chosen because of their potential benefits to Spotify. By attempting to predict the characteristics that make a song popular on the platform and which characteristics comprise hit songs, Spotify can fine-tune its recommendation algorithms to include more popular songs in users' playlists. Automatic Playlist Continuation (APC) benefits users' ability to listen to and create playlists by allowing them to extend their listening experience beyond an existing playlist's end and create compelling playlists more easily (Zamani *et al.*, 2019). Additionally, by determining whether specific features contribute to a song's popularity or its potential to become a hit on the Billboard Hot 100 Chart, the company can discover if there is a particular type of song they should be marketing more heavily or including more often on their platform.

The author used the Wilcoxon rank-sum test (WRST) and the Kruskal-Wallis (KW) test on both datasets to answer the first business question. These two nonparametric tests determine the differences between groups when the populations do not meet the assumptions required by the two-sample *t*-test and the analysis of variance (Lind *et al.*, 2021). The author used Dunn's test to determine which groups are significantly different from each other when rejecting the KW test results (Dinno, 2015). Multiple linear regression was used on both datasets to determine how successfully one could predict a song's popularity. Finally, the researcher used SAS Enterprise Miner to analyze output from various models to determine how successfully one could predict a song's chances of reaching the number one spot on the Billboard Hot 100 Charts. The models compared include logistic regression models, decision trees, neural networks, rule induction models, and ensembles.

Multiple logistic regression describes a relationship between several predictor variables and a dichotomous criterion variable. Decision trees use a series of "if-then-else" rules to generate a predicted value. Neural networks attempt to emulate a neuron's behavior, using multiple layers of nodes to predict a criterion variable. Rule induction models use "if-then-else" rules to improve the classification of rare events. Finally, ensembles combine two or more predictions from predictive models into a single composite score (Abbott, 2014).

A sample node was used to balance the dataset, ensuring an equal number of hits and non-hits. Additionally, the researcher created a target profile using the appropriate prior probabilities to ensure each model was tested using the criterion variable's original proportions, not the proportions in the oversampled data. The data partitioning node divided the data into training and validation sets, each containing 70% and 30% of the data, respectively. Several metrics that seek a balance between false positive and false negative rates, including the geometric mean, discriminant power, and balanced accuracy were calculated in Excel to evaluate these models (Akosa, 2017). The fact that Spotify is an international organization did not affect the dataset, variables, or tools and techniques needed to analyze the data, as these tools are available on publicly accessible websites.

Limitations

The study's limitations include that it was constrained to an 8-week time frame and that the researcher performed a secondary analysis on two pre-existing datasets. In such a scenario, it is up to the researcher to determine the data's credibility. The dataset obtained from Kaggle.com has been downloaded from the website 17,200 times and utilized by 67 publicly available notebooks (*Spotify Dataset 1921-2020, 160k+ Tracks*, n.d.). Additionally, the dataset contained many fields used in other published research studies, including those conducted by Al-Beitawi *et* *al.* (2020), Herremans *et al.* (2014), and Ni *et al.* (2011). Moreover, the researcher used two datasets to validate his results, though one of the datasets was imbalanced. The student used oversampling to overcome this challenge, as described above. Therefore, the author thought the analysis of these two datasets offered valuable insights into HSS. The fact that the author picked a non-U.S. organization for his research project did not provide additional challenges to his analysis.

Ethical Considerations

Boté & Térmens (2019) write that several ethical concerns arise when considering the reuse of data, including data trustworthiness, informed consent, and data anonymity. This study's dataset contained no identifying information, therefore alleviating the researcher of ethical considerations regarding informed consent and data anonymity. Additionally, there are currently no standards to validate a dataset's level of trustworthiness regarding data reuse, and researchers often rely on their own evaluations of the data to determine trustworthiness. To overcome any additional ethical challenges, the researcher adhered to all terms and conditions posted on relevant websites ("Community Data License Agreement – Sharing, Version 1.0," n.d.; *Spotify Developer Terms of Service | Spotify for Developers*, n.d.; *Terms of Use*, n.d.). The fact that Spotify is an organization outside of the U.S. did not affect security, privacy, or ethical concerns.

Findings

The results of the WRST performed on the first dataset comparing song popularity distributions grouped by song mode indicate that songs deriving their melodic content from the minor scale are significantly more popular on Spotify than songs deriving their melodic content from the major scale ($\chi = 148.53$; $p < .0001$). The same analysis performed on the second dataset confirms these results ($\chi =312.45$; $p < .0001$). Next, the results of the WRST performed on the first dataset comparing Spotify song popularity distributions grouped by explicit lyrical content indicate that songs with explicit lyrical content are significantly more popular on Spotify than songs without this content ($\chi = 6211.08$; $p < .0001$). The same test performed on the second dataset confirmed these results ($\chi = 2549.32$; $p < .0001$). The KW test results comparing song popularity distributions by song key for the first dataset indicate significant differences amongst these groups ($\chi = 1608.7196$; $p < .0001$). Dunn's test indicates that C# is significantly more popular than all other keys besides F# and B. The same analysis performed on the second dataset indicates similar findings ($\chi = 392.39$; $p < .0001$). Once again, Dunn's test confirmed that C#, F#, and B are the most popular keys on the Spotify Platform. Figures 1 – 6 display these results.

The multiple linear regression analysis performed on the first dataset indicates a single variable, namely the year, explained 74.38% of the variation in a song's popularity score. However, after removing this variable from the dataset, the best model explained less than 50% of song popularity variation, indicating substandard prediction capabilities. The same analysis performed on the second dataset confirmed these results. After removing the week that a song appeared on the Billboard charts, the best predictive model explained just 32.10% of song popularity variation ($F = 795.60.4$; $p < .0001$). Figures 7 – 9 display these results. Figures 14 – 17 display the Base SAS code to perform these analyses, which the author uploaded to his GitHub account (Miner, 2021).

Finally, the author used SAS Enterprise Miner and Excel to determine how successfully one could predict a song's chances of becoming a hit on the Billboard Hot 100 Charts. SAS Enterprise Miner was used to build, train, and test the models, while Excel was used to calculate each model's accuracy, misclassification rate, sensitivity, specificity, and precision. Table 2 provides these results. Figures 10 – 13 show the analyses performed in SAS Enterprise Miner and the program's output.

However, Akosa (2017) writes that predictive accuracy can be a misleading measure of model performance when datasets are imbalanced because these measures assign more weight to the majority class than the minority class. Therefore, analysts should seek measures that balance false positive and false negative rates. Akosa lists several of these measures, including the geometric mean, discriminant power, balanced accuracy, and Youden's Index, which provide more robust performance metrics for imbalanced datasets. The author calculated these metrics in Excel. Table 3 provides these results, showing that logistic regression with a backward selection technique outperformed all other models. Figure 12 shows this model's results and that *acousticness, danceability, liveness, duration*, and *explicitness* correlate significantly with the criterion variable. For each unit increase in *acousticness* and *danceability*, the odds of a song becoming a hit are estimated to be 2.99 and 5.41 times greater, respectively, whereas for each unit increase in *liveness*, the odds of a song becoming a hit decrease by a factor of 0.27. If a song contains explicit lyrics, its chances of becoming a hit are decreased by a factor of 0.62.

Akosa (2017) writes that the discriminant power metric assesses how well a classifier distinguishes between positive and negative classes. A classifier is considered substandard if its discriminant power is below one. Logistic regression with a backward selection technique obtained the highest discriminant power at 0.22, but this measure is far below one. Therefore, we once again reject the null hypothesis and conclude there are statistically significant relationships between a song's chances of achieving a number one spot on the Billboard Hot 100

Chart and other song characteristics. However, the discriminant power metric shows the model's predictive power is relatively weak.

Conclusion

In conclusion, this research demonstrated some evidence for Hit Song Science. For instance, we saw the distributions of popularity scores are not the same for songs grouped by song mode, explicit lyrical content, and song key. We discovered that songs deriving their melodic content from the minor mode are significantly more popular on Spotify than songs deriving their melodic content from the major mode and that songs containing explicit lyrical content are significantly more popular than songs without this content. Moreover, we saw that songs in the key of C#, F#, and B are significantly more popular on Spotify than songs written in other keys. We discovered that many features in the dataset were significantly correlated with predicting a song's popularity and its chances of becoming a hit on the Billboard Hot 100 Charts, though our models showed weak predictive capabilities. Finally, we saw that logistic regression using a backward selection technique outperformed all other models in determining a song's chances of becoming a success on the Billboard Hot 100 Charts.

Recommendations

Future research recommendations include determining whether it is possible to predict a song's chances of becoming a top 5 or top 10 success rather than a top 1 hit, performing analyses according to specific genres, and adding lyrical content to the dataset. By attempting these future analyses, we may improve the accuracy of predictive models and identify songs that are potential successes more precisely. In turn, Spotify could more efficiently tailor its recommendation algorithms and targeted marketing campaigns to increase profits.

References

Akosa, J. (2017). *Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data*. 12.

Al-Beitawi, Z., Salehan, M., & Zhang, S. (2020). What Makes a Song Trend? Cluster Analysis of Musical Attributes for Spotify Top Trending Songs. *Journal of Marketing Development & Competitiveness*, *14*(3), 79–91.

Boté, J.-J., & Térmens, M. (2019). Reusing Data: Technical and Ethical Challenges. *DESIDOC Journal of Library & Information Technology*, *39*(6), 329–337. https://doi.org/10.14429/djlit.39.6.14807

Community Data License Agreement – Sharing, Version 1.0. (n.d.). *CDLA*. Retrieved January 8, 2021, from https://cdla.dev/sharing-1-0/

Dhanaraj, R., & Logan, B. (2005). Automatic prediction of hit songs. *In Proceedings of the International Conference on Music Information Retrieval (ISMIR*, 488–491.

Dinno, A. (2015). Nonparametric Pairwise Multiple Comparisons in Independent Groups using Dunn's Test. *The Stata Journal*, *15*(1), 292–300. https://doi.org/10.1177/1536867X1501500117

Eltringham, J. (2017). Archiving Spotify: Methods and Motives for Collecting Individual Music Streaming Data. *Journal of Archival Organization*, *14*(3/4), 128–138. https://doi.org/10.1080/15332748.2018.1505386

Fielding, A. (2020, February 10). *Spotify reports total revenue of $7.44 billion in 2019*. DJMag.Com. https://djmag.com/news/spotify-reports-total-revenue-744-billion-2019

Fitzgerald, S. M., Rumrill, P. D., Jr., & Schenker, J. D. (2004). Correlational designs in rehabilitation research. *Journal of Vocational Rehabilitation*, *20*(2), 143–150.

Fleischer, R., & Snickars, P. (2017). Discovering Spotify-A thematic introduction. *Culture Unbound*, *9*(2), 130–145.

Haupt, J., & Shelley, A. (2012). Spotify. *Notes*, *69*(1), 132–138.

Herremans, D., Martens, D., & Sörensen, K. (2014). Dance Hit Song Prediction. *Journal of New Music Research*, *43*(3), 291–302. https://doi.org/10.1080/09298215.2014.881888

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2021). *Statistical techniques in business & economics* (Eighteenth edition). McGraw-Hill Education.

Marshall, L. (2015). 'Let's keep music special. F—Spotify': On-demand streaming and the controversy over artist royalties. *Creative Industries Journal*, *8*(2), 177–189. https://doi.org/10.1080/17510694.2015.1096618

Miner, S. (2021). *Sminerport/CapstoneProject* [HTML]. https://github.com/sminerport/CapstoneProject (Original work published 2021)

Miller, S. (n.d.). *Billboard Hot weekly charts—Dataset by kcmillersean*. Data.World. Retrieved January 23, 2021, from https://data.world/kcmillersean/billboard-hot-100-1958-2017

Narayanan, A., Shi, E., & Rubinstein, B. I. P. (2011). Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge. *ArXiv:1102.4374 [Cs]*. http://arxiv.org/abs/1102.4374

Ni, Y., Santos-Rodriguez, R., Mcvicar, M., & De Bie, T. (2011). Hit song science once again a science. *4th International Workshop on Machine Learning and Music*.

O'Leary, Z. (2017). *The essential guide to doing your research project* (3rd edition). SAGE Publications.

Pachet, F., & Roy, P. (2008). *Hit Song Science Is Not Yet a Science.* 355–360.

Spotify Dataset 1921-2020, 160k+ Tracks. (n.d.). Retrieved January 23, 2021, from https://kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

Spotify Developer Terms of Service | Spotify for Developers. (n.d.). Retrieved January 8, 2021, from https://developer.spotify.com/terms/

Spotify employees count 2018. (n.d.). Statista. Retrieved January 23, 2021, from https://www.statista.com/statistics/245130/number-of-spotify-employees/

Spotify—Company Info. (2020, September 30). Spotify. https://newsroom.spotify.com/company-info/

Terms of Use. (n.d.). Data.World. Retrieved January 24, 2021, from https://data.world/policy/terms/

Wikhamn, B. R., & Knights, D. (2016). Associations for Disruptiveness—The Pirate Bay vs. Spotify. *Journal of Technology Management & Innovation*, *11*(3), 40–49. https://doi.org/10.4067/S0718-27242016000300005

Zamani, H., Schedl, M., Lamere, P., & Chen, C.-W. (2019). An Analysis of Approaches Taken in the ACM RecSys Challenge 2018 for Automatic Music Playlist Continuation. *ArXiv:1810.01520 [Cs]*. http://arxiv.org/abs/1810.01520