OSCARS PREDICTION MODEL

-Final Project Report

Adithya A S Department of Computer Science and Engineering PES University *Bengaluru,India* adithyaaravi10@gmail.com

Chirag Rudresh Department of Computer Science and Engineering PES University *Bengaluru,India* chiragrudresh15@gmail.com

P Deepak Reddy Department of Computer Science and Engineering PES University *Bengaluru,India* deepakreddy14@gmail.com

***Abstract*—The paper includes an in-depth analysis of the problem at hand - Oscar winning predictions. The issues faced in modern times with respect to the topic and the ideology behind creating a predictive model is described in the paper.The paper also gives overall review of the existing solutions.It then goes on to describe the working of the model and the need for such a model in the competitive world of movie media.**

# *I. Introduction*

Sitting on the couch on a lazy Sunday afternoon, you decide to catch up on the latest movie buzz, maybe go through some of the movies that sparked the Oscar talk. As you scroll through the vast list of movies released this year and not to mention the ones yet to be released, a thought wanders your mind - How does one choose a single movie among the thousands and millions released worldwide to be the winner of Best Movie Oscar?

Movies have been one of the most sought after forms of entertainment for as long as time can remember. Not only is the industry a phenomenal business but the fame and history that goes along every little detail of a movie can make or break the cast’s career. For a common civilian, it was just a weekend theatre time for time pass or for a critic it was a tedious job, but for the movie industry every Friday release is as important as betting every penny of their life’s earnings on the movie success. How a film does, not only in the box office, but also among the trend-seeking crowd impacts the lives of many people, from the lead director to the spot boy.

Out of the millions of movies and short films released worldwide every year, the enormous responsibility to select the best of the best for every possible category in the movie industry is given to a special set of board members called the Oscar Awards Committee. These few men and women decide the fate of the most sought after, once in a lifetime prestigious Oscars. The task is a never ending, year round job that requires attention to detail and a sound knowledge of the movies. It might just be an award show night for the viewers, but there is a lot at stake for the filmmakers and their reputations. While some directors and scriptwriters make films for the audience to have a lasting memory, some strive to achieve their childhood dream of etching their name on the Academy Award trophy forever.

The Oscar winnings are so important that it has become some of the filmmakers primary goal to make movies for the sole purpose of getting nominated for the Awards, let alone win one. The competition, as we know, is seemingly endless with every Friday bringing upto 600 movies being released in the United States alone. The estimation for what can win an Oscar in the earlier days, were just made by well known critics based on their intuition, often biased due to personal sentiments. But with the fast growing technology of analytics and the abundance of data available on every possible topic, the world of possibilities has just become endless. Needless to say this technology is a boon and a bane for many filmmakers as this allows one to estimate their movie winning or being nominated even more clearly but also increases the competition ten folds.

The model reported in the study, seeks to combine technology of data and analytics with the ever growing movie industry to predict what are the chances of a nominated film to win an Oscar in any category.

*II.Previous Work*

[1]The models used previously involved data regarding only Movie titles and few features such as Votes or Ratings and so on. Many models use a selective majority voting Decision Tree or an Artificial Neural Network to predict the probability based on such fears extracted from given data. Most of the previous work involves analysis of selected data and representing the results of trend- analysis as graphs and thus making assumptions manually. [2]These models also aim to predict the winning percentage in general for movies and not specifically for a single movie.Many papers have used a completely different approach where they have used sentimental analysis techniques to understand how a movie can win an Oscar based on people’s opinions and discussions.

Moreover, these models predict based on select features such as only for a particular year or based only on IMDb ratings from users. Thus not encapsulating the whole idea of how Oscars are decided every year.Other analysis on this topic included a simple survey of public .[3]The limitations with these models and analysis is that they cherry pick the data features and base the probability of winning on a single or very few data features.

From the data set we have we begin to analyse the different attributes present and based on general intuition make assumptions on how these attributes(like user ratings,duration) can be related to Oscar winnings.For instance our model assumes that the other achievements of a movie throughout the year doesn’t effect the chance of it winning an Oscar, so we considered summing up all the achievements onto a single column .We also assumed that the nominations received for the film in other awards affected the predictions and thus we considered that to be an important feature in our prediction model. Ideally there are many details that go into deciding the winners of Oscar and our model uses this as an example to set more features and use them productively to make a better and more accurate model.We implement this concept using a statistical approach to understand the data and to find out the optimal features that increase the probability of a movie winning an Oscar.

# *III.Proposed Solution*

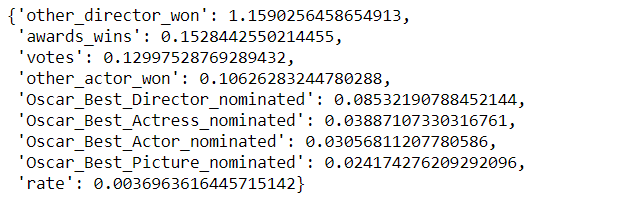
Our goal was to predict the probability of a movie winning Oscars in the category of Best Pictures,Best Director, Best Actor and Best Actress and in-order to achieve this it was important to understand all the attributes present in our dataset and to extract the required features.We started off with the important step of pre-processing the data at hand. Each attribute was carefully analysed on the number of null values present ,type of the data dealing with and how it was related to other attributes in the column. Appropriate methods were applied to each column to fill null values and to handle outliers.After many careful considerations it was understood that the model was performing poorly when many attributes were used so as a part of the dimensionality reduction process new columns were generated to replace the existing attributes.With the new attributes in hand the data set was split into train and test with training data containing values from 2001 to 2016 and test data containing values of 2017.Since we were dealing with four different kinds of Oscar categories and each category had different features affecting the chances of winning an Oscar, we identified these features for each category with the help of Linear regression model and found out the percentage of importance (i.e how useful an attribute is important in prediction the final output) for each attribute and identified a list of attributes which gave positive value.Considering only these respective attributes, initially correlation heatmap was built to visualize how each attribute were dependent on each other .Later scatter plot was visualised with the output label and the highest important attribute for each category and it was found that these plots looked similar.After this individual Linear Regression models were built for each category using only the list of important attributes and the model predicted probability for a movie to win an Oscar in the respective category.Considering the movies with top five probabilities for each category ,it was found that one of those five movies had won an Oscar for that particular category in the year 2017.

## *[1]Preprocessing*

The initial data set contained 119 columns which included categorical,numerical and text data having outliers and null values. Each column was analysed individually and as per the requirement pre-processing was done.For instance,the actual date of movie release was looked up on the internet and filled inplace of ‘Year’ null values.The ‘Gross Income’ null values were filled by taking only the average of other movies in that particular year since it is likely that the overall income of movie can change annually.Since the correlation between ‘User Reviews’ and ‘Popularity’ was high, the best fit line was found between these two columns and null values were filled accordingly.Similar process was also used to fill null values for ‘Critic Reviews’ by finding best fit line with ‘User Reviews’.As stated in our list of assumptions, we had assumed that individual awards do not affect the overall Oscar wins,so in order to reduce the number of columns we summed up the total number of non Oscar wins and nominations onto two columns called ‘Award Wins’ and ‘Award Nominations’.All the categorical data were converted to numerical data.In the case of ‘Certificate’ total number of distinct elements were found and numerical values were substituted according to their weights.For instance,‘Unrestricted’ comes first which is followed by ‘U/A’ according to the Central Board of Film Certification,hence these two were replaced with 1 and 2 respectively since these attributes are close to each other.For Oscar awards and nominations all the ‘Yes’,’No’ values were replaced with 1 and 0 respectively.It was also necessary to find the count of awards and nominations the movie had won other than Oscars for each category.This was done with the help of text processing on the ‘Award Categories’ column.For instance the string ‘Best Director’ was found to be the most common for every movie that won ‘Golden Globes for Best Director ’ award .Similarly all the common strings for Best Picture, Actor and Actress were looked up manually .Simulation containing these strings was run later to extract count of wins for each categories for Golden Globes.The same process was repeated for both award wins and nominations of awards like BAFTA,Critics Choice,People’s Choice and many other awards to find the count of award wins and nominations for individual category.After preprocessing, the entire data set was reduced to 28 columns with no null values.

## *[2]Model Building*

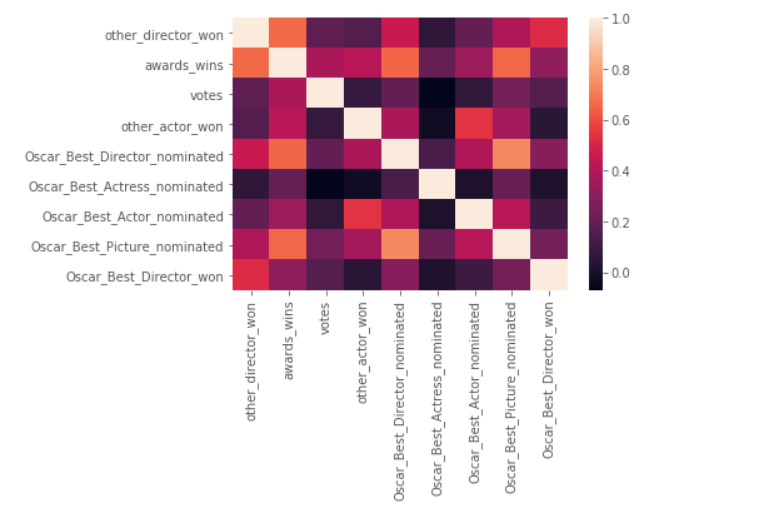
The next step was to choose the set of attribute that would help in predicting the probability of winning for each category.There were many available regression models and after analysis from online sources, Linear Regression was chosen since it provided better results compared to other models.The data set was now split based on years since we aim to train the model with the data from 2001-2016 and to predict the values of 2017.The entire test train data set was normalised so that no individual would intrinsically influence the result more due to its larger value and to bring all variables under the same range.Next step was to calculate the ‘Importance’ of each attribute for a category with the help of Linear Regression.After getting these values only attributes with positive values were taken into consideration since those play a major role in deciding if a movie would win Oscar or not.

*Fig 1.1 - Importance of each attribute in Director*

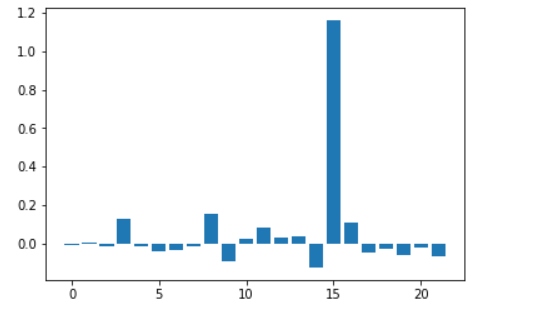
The important attributes varied for other categories. Later these were used to train the Linear Regression model.After training the model test data containing values values from 2017 were passed and predictions were accounted for.

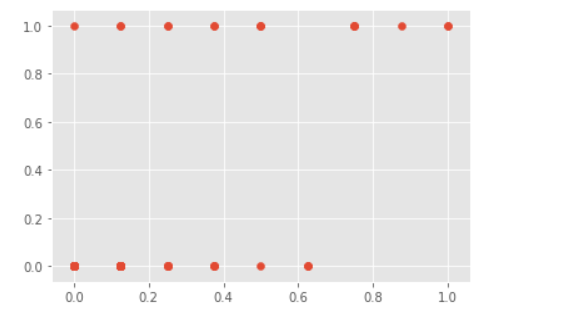
## *[3]Evaluation*

Softmax function was applied to the top 5 normalised prediction values to find probability of winning of these movies and was later checked on the internet to see if they won an Oscar or not.The important attributes were also plotted along the correlation heatmaps

*Figure 2.1 - Heatmaps between important attributes of director*

These important attributes were found by plotting bar graphs along with the total set of attributes available

*Figure 2.2 - Bar graph of total set of attributes for Director*

It was found that the scatter plot for the maximum important attribute with the output label was similar through all categories 

### *Figure 2.3 - Scatter Plot between Actor and its Most Important attribute*

### *Figure 2.4 - Scatter Plot between Director its Most Important attrib*

# *IV.Experimental Results*

Using the above model we have predicted the 5 most likely movies to win the Oscars in each of the four categories

Here are our predictions for 2017 Oscars.

| **Rank** | **Movie** | **Probability** |
| --- | --- | --- |
| 1 | The Shape of Water (actual winner) | 0.343 |
| 2 | Dunkirk | 0.180 |
| 3 | The Florida Project | 0.160 |
| 4 | Call Me by Your Name | 0.158 |
| 5 | Darkest Hour | 0.156 |

## *Table 1.1 -Predictions for Best Director*

These are the top 5 likely movies to win the Oscar for the category ‘Best Director’ . The actual winner is ‘Shape of Water’ , which we predicted as the most likely movie to win.

| **Rank** | **Movie** | **Probability** |
| --- | --- | --- |
| 1 | Get Out | 0.250 |
| 2 | Lady Bird | 0.215 |
| 3 | Three Billboards Outside Ebbing, Missouri | 0.192 |
| 4 | Call Me by Your Name | 0.171 |
| 5 | The Shape of Water (actual winner) | 0.169 |

## *Table 1.2 -Predictions for Best Picture*

These are the top 5 likely movies to win the Oscar for the category ‘Best Picture’ . The actual winner is ‘Shape of Water’ , which we predicted as the fifth most likely movie to win.

| **Rank** | **Movie** | **Probability** |
| --- | --- | --- |
| 1 | Call Me by Your Name | 0.316 |
| 2 | Darkest Hour  (actual winner) | 0.256 |
| 3 | Get Out | 0.161 |
| 4 | The Post | 0.132 |
| 5 | Logan | 0.132 |

## *Table 1.3 -Predictions for Best Acto*r

These are the top 5 likely movies to win the Oscar for the category ‘Best Actor’ . The actual winner is Gary Oldman for the movie ‘Darkest Hour’ , which we predicted as the second most likely movie to win

| **Rank** | **Movie** | **Probability** |
| --- | --- | --- |
| 1 | The Shape of Water | 0.249 |
| 2 | Lady Bird | 0.217 |
| 3 | Three Billboards Outside Ebbing,Missouri  (actual winner) | 0.207 |
| 4 | I , Tonya | 0.167 |
| 5 | The Post | 0.158 |

## *Table 1.4 -Predictions for Best Actress*

These are the top 5 likely movies to win the Oscar for the category ‘Best Actress’ . The actual winner is Gary McDormand for the movie ‘Three Billboards Outside Ebbing,Missouri ’ , which we predicted as the third most likely movie to win.

The possibility of a movie winning an Oscar in a respective category , can very well be determined on the basis of the number of nominations and awards it has achieved in other film awards rather than how well the audience has received it .This trend is reflected in the years 2001 - 2016 . Ratings and reviews/ratings by audience only may not be a good factor to determine the possibility of winning an Oscar. This model worked well even with less data , the dataset used in this model is for only 17 years (2001-2017) with only 17 positive samples against around 1000 negative samples.

The model fails to take into account the unforeseen factors, for example, when Leonardo DiCaprio won the Oscar for lead actor category for the film ‘The Revenant’ , critics claimed this to be debatable, Everyone suggests that DiCaprio had missed out too many times before and that now he should be rewarded.It may also fail in the cases where in a particular year for which Oscar winners are to be predicted , there isn’t any standout film and many movies have won various film awards.Model may not be accurate if necessary data isn’t available .This happens in cases where other film awards are not hosted before the Oscars

# *V.Conclusion*

Each member of the team was an integral part of the research and implementation of this prediction model. The building of this model were divided in to the following steps :

Research and Problem finding : All three members of the team put thorough effort into finding a problem statement and analysing the need for type of data and a suitable model to solve the problem.

Data collection : Chirag Rudresh aided in finding the final data set from Kaggle and thus the team further analysed the dataset and fine tuned the columns required.

Pre-Processing : Adithya AS and Chirag Rudresh cleaned the data, which involved detecting and removing null values and outliers. Deepak Reddy found correlations between columns in data which helped in solving the prediction problem. The team divided the dataset of 119 columns into 3 equal parts and performed further preprocessing that was removing rows if necessary and converting categorical data to binary.

Building the Model and Testing. : The whole team was involved in the process of building the model from scratch and discussing every detail including parameters and debugging.

Final Report and Videos : The task of writing the final report was split equally among the three members each having to write 2 pages approximately. The video recording and presentation were also a joint effort by the team members.

# VI. References

[[1] Walt Hickey , ‘FiveThirtyEight’s Guide To Predicting The Oscars’ , Jan 2016](https://fivethirtyeight.com/features/fivethirtyeights-guide-to-predicting-the-oscars/)

[[2]Nicholas Parker, ‘Predicting the Oscars using Preferential Machine Learning’ , Feb 2020](https://towardsdatascience.com/predicting-the-oscars-using-preferential-machine-learning-32f06ffbf427)

[[3]Aatkan Cetinsoy, ‘Predicting the 2017 Oscar Winners With BigML’ , Feb 2017](https://dzone.com/articles/predicting-the-2017-oscar-winners)