TMDB Analysis

Ameeshi Gupta | Darren Chan | Koh Jia Cheng

AGENDA

Introduction

Problem Statement

Wrangling

Data Collection & Organisation

Exploratory Data Analysis

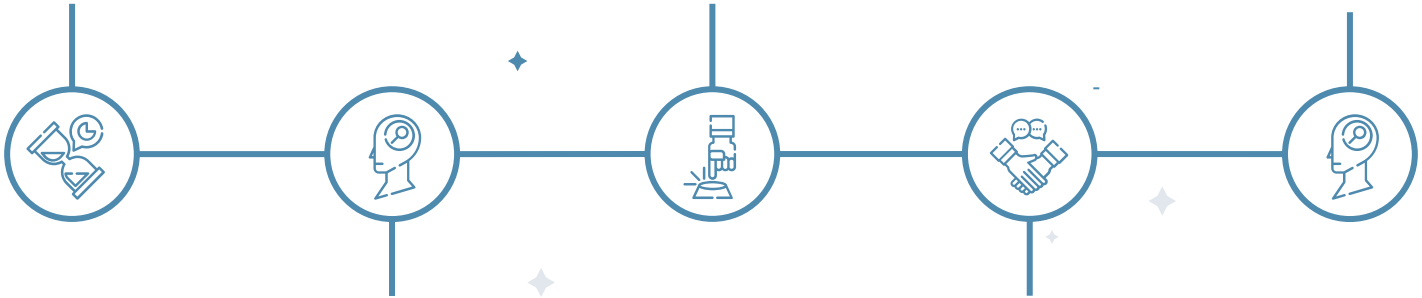
Analysis & Observations

Machine Learning

Logistics Regression Model

Conclusion

Outcome & Insights



Problem Statement

Will The Film Be Successful?

Film's investor point of view to determine if the production will be profitable & what factors contribute to a film's success

DATA WRANGLING

API Call

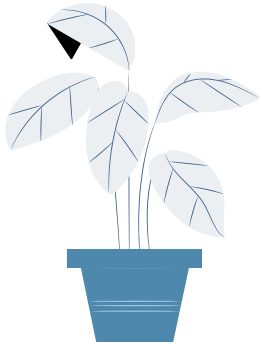
Extract and Store in a CSV file

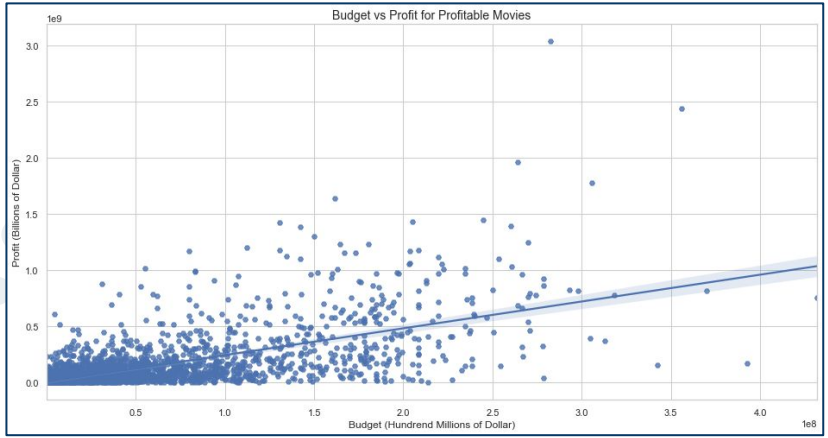
Real Value

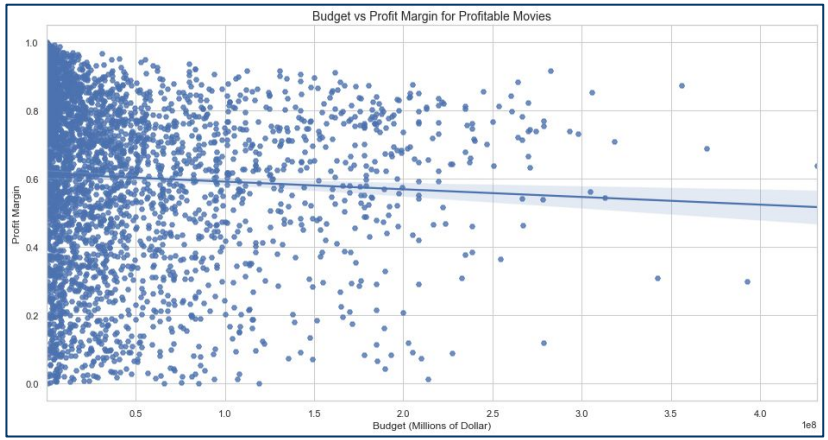
Using US Inflation Rate

RegEx & Inline

Clean & Group the data

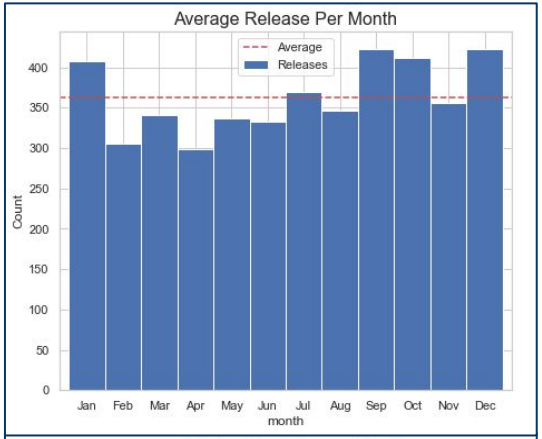


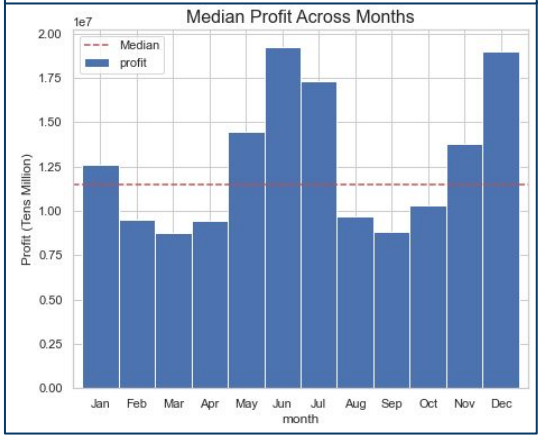




Budget vs Profit Margin

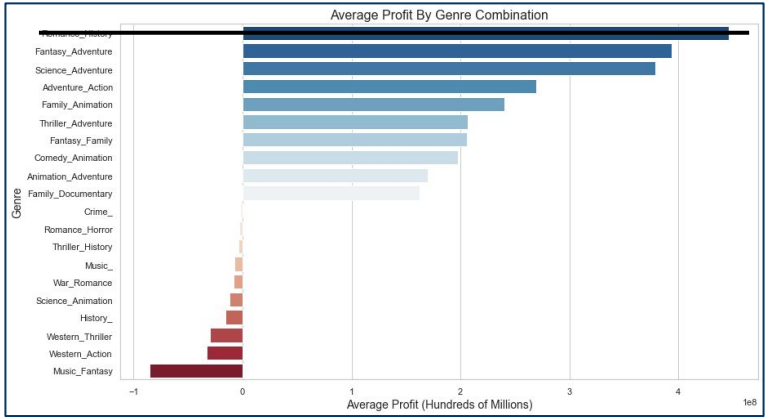
- Negative Correlation with *Budget* and *Profit Margin*
- Diminishing Profit Margin as *Budget* increases
- *Budget* should be kept to a minimum

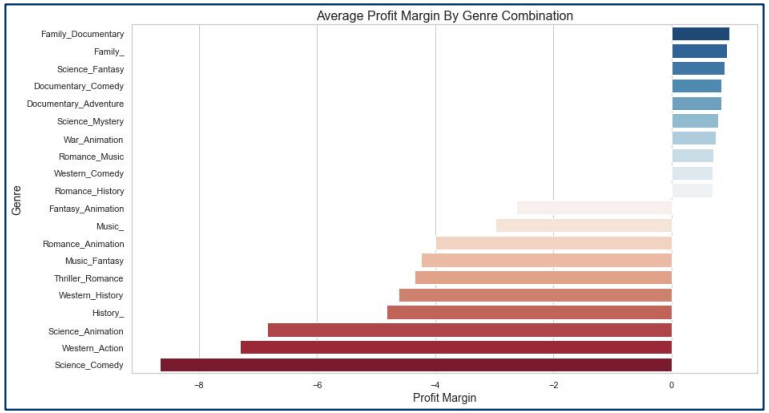




Profit by Months

- September to January less November have the highest release of films
- June, July & December are months that have a high *Profit*



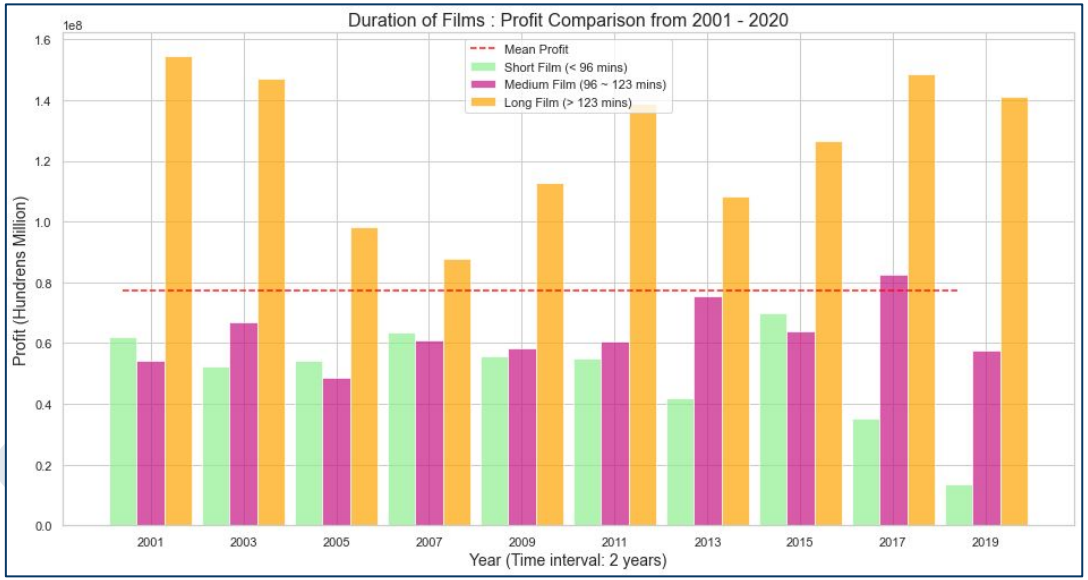


Average Profit by Genre

- ➔ Dual Genres generally fare better than single *Genre*
- ➔ High *Profit* does not translate into high *Profit Margin*
- ➔ Documentary would be a safe choice for profit margin

```
Number of Films Released in the Top Genres Released:
No. Romance History Film 1
No. Fantasy Adventure Film 66
No. Science Adventure Film 9
```

Profit by Films' Duration



- Films longer than >123 mins fetch higher *Profits*
- This phenomenon can be observed since early 2000s, having a parabola effect to 2019

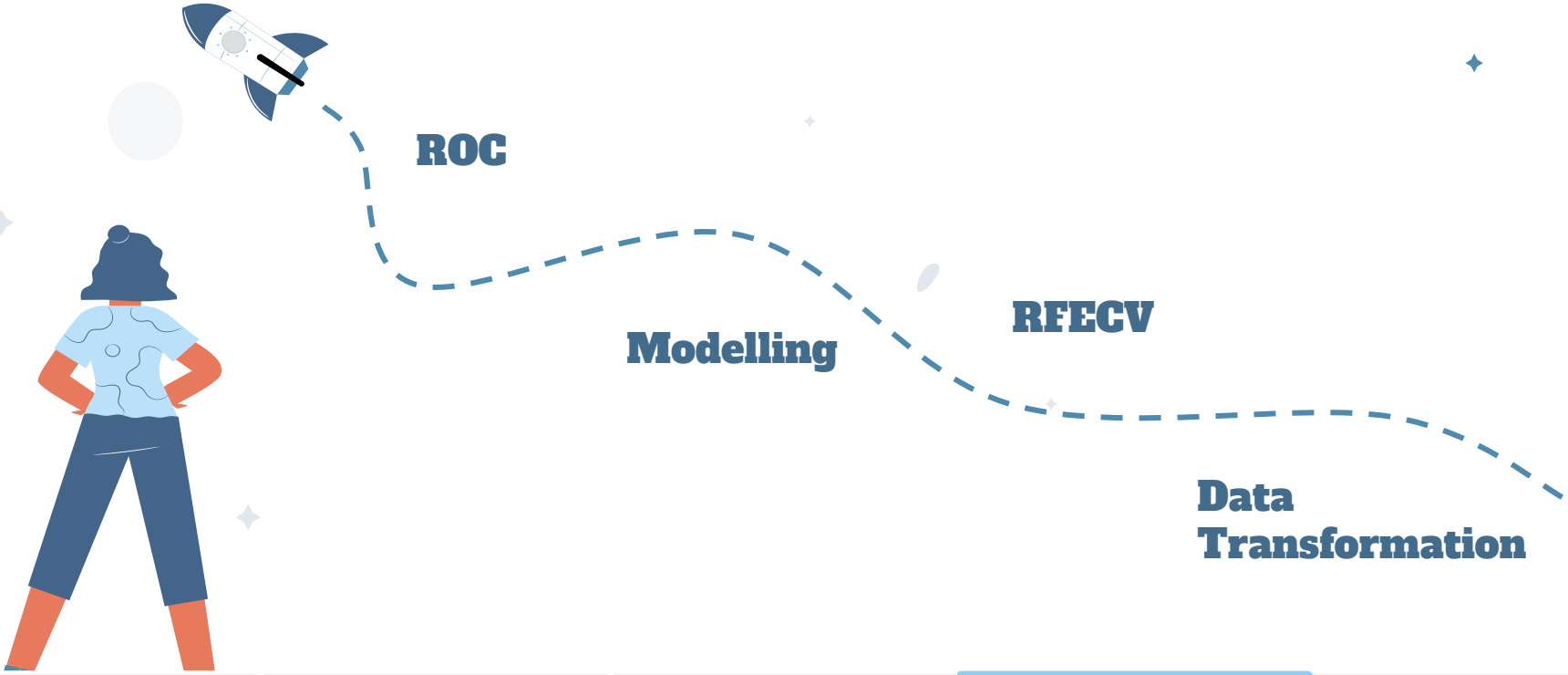
Machine Learning

Logistics Regression to predict whether a movie will be profitable based on attributes that are only available prior to it's release.

The Attributes:

- Budget, Popularity,
- Avg_Cast_Var,
- Director_Var,
- Production_Var,
- Day_Of_Week,
- Month,
- genre_combo,
- Run Time

Implementation Steps



Data Preparation

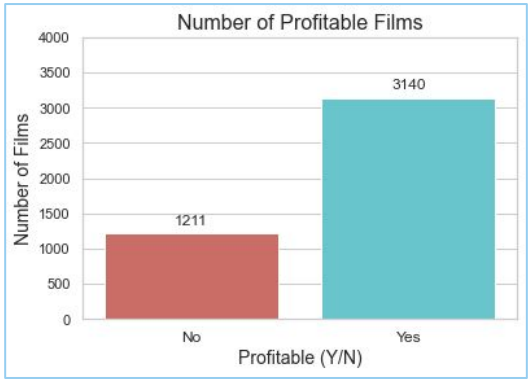
1. Generate Dummy Variables

   Reference Level:
   day_of_week_fri, month_apr, genre_combo_Action_

2. Train-Test-Split with Stratify

3. Standardisation

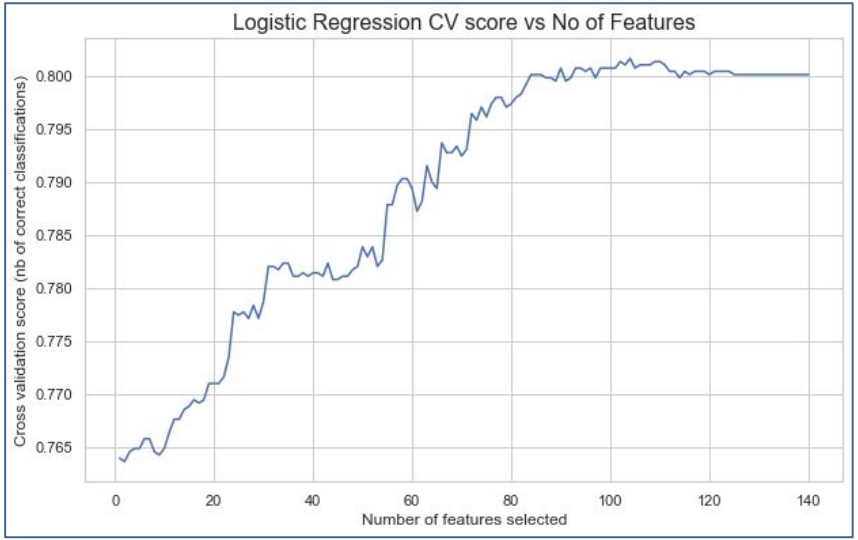
month_Feb	month_Jan	month_Jul
0	1	0
0	0	0




```
# Sanity Check
X_train.std()
```

```
budget          1.000153
popularity      1.000153
avg_cast_var    1.000153
```

Recursive Feature Elimination Cross Validation




```
from sklearn.feature_selection import RFE, RFECV
from sklearn.linear_model import LogisticRegression
#Logistic Regression Model
logreg_model = LogisticRegression(class_weight = 'balanced', random_state = 0)
rfecv = RFECV(estimator=logreg_model, step= 1, cv = 5, scoring='accuracy')
rfecv = rfecv.fit(X_train, y_train)
```

Before

```
Train Shape: (3263, 140)
Test Shape: (1088, 140)
```

RFECV

Optimal Features: 104

Fit and Model

The Model is 0.96 right in predicting Profitable Film, which is what we want.

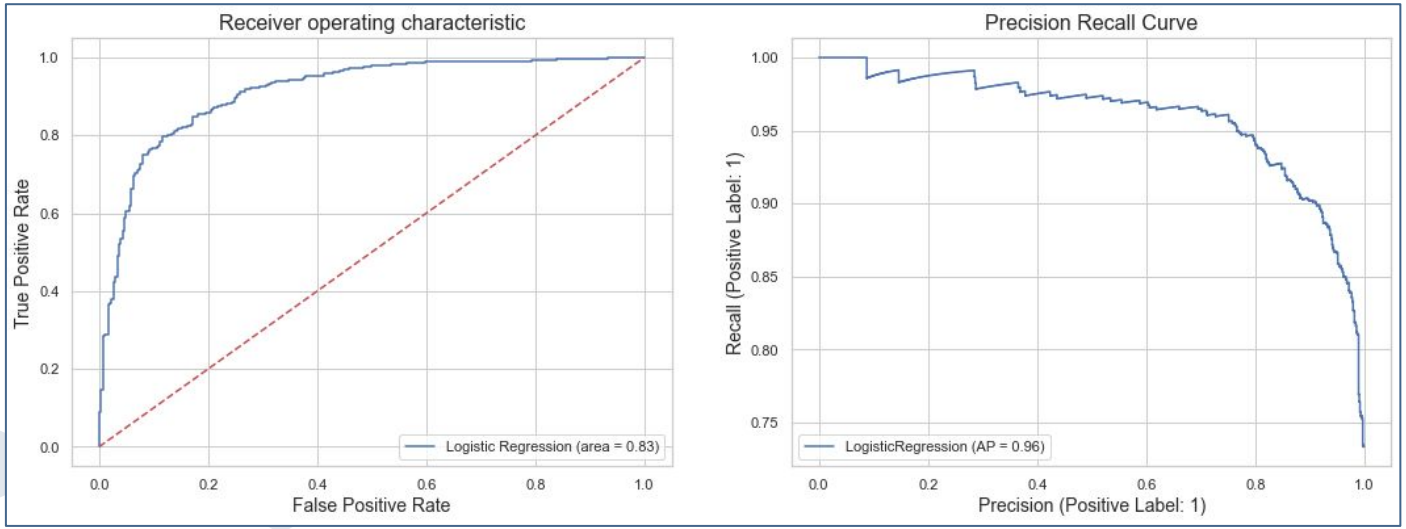
Confusion Matrix

Classification Report

	Precision	Recall	FI-score	Support
0	0.59	0.92	0.72	303
1	0.96	0.75	0.84	785
Accuracy			0.80	1088
Macro Avg	0.77	0.83	0.78	1088
Weighted Avg	0.86	0.80	0.81	1088

Opportunity Lost is better than Money Lost!

ROC and Recall Curve



Good Precision and Recall

Interpreting Result

	Variable	Coefficient	
0	budget	-0.853982	Exp(-0.85) -> 0.42
1	popularity	-0.299806	
2	avg_cast_var	0.532002	Exp(0.53) -> 1.69
3	director_var	0.438631	
4	production_var	7.693367	
5	runtime	0.280174	
8	day_of_week_Thu	0.189285	
9	day_of_week_Tue	-0.221630	
10	month_Aug	-0.198156	
11	month_Dec	0.129855	
35	genre_combo_Family_Comedy	1.055084	Exp(1.055) -> 2.87
42	genre_combo_Fantasy_Family	0.835923	
74	genre_combo_Science_Adventure	0.797879	
73	genre_combo_Science_Action	0.795221	

Explanation

1. For Categorical, odds ratio is to a referenced Categorical Data

2. Numerical is in terms of SD, odds to increase Profitability.

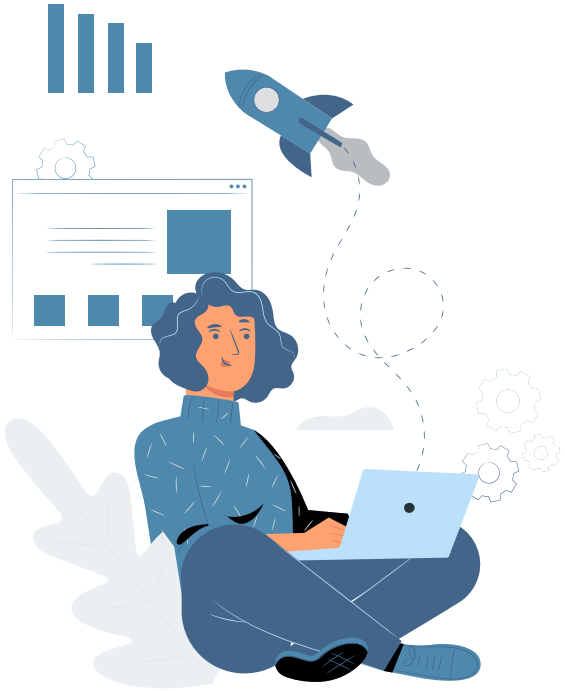
3. Matches with EDA insights

Conclusion

Problem Solved - Key Insights from EDA & Modelling

- Production Studio, Director & Cast are Important for Profitability (1 out of top 5)
- Higher Runtime has a Higher Probability of Profitability (>123 min)
- Higher Budget does not Translate to Higher Profitability
- Science & Adventure and Family & Comedy genres are expected to be profitable
- Release Date can be Dec for higher Profitability.

Additional Learnings



1. Regular Expression
2. Dummy Variable/ One-Hot Encoding
3. Receiver Operating Characteristic (ROC) Curve
4. Recursive Feature Elimination Cross Validation (RFECV)
5. Importance of Cross-Validation
6. Decoding JSON format during the use of an API

Workload Distribution

Ameeshi

- Analysis Production VARs
- Editing of slides
- Filming

Darren

- Runtime Analysis
- Popularity Analysis
- Machine Learning
- Reference Material
- Slides, Visuals and Filming

Jia Cheng

- Data Extraction, Cleaning, Preparation
- Rest of EDA
- Machine Learning
- Slides and Filming

Please checkout our notebook too, there's additional sub questions there.

References

1. https://stackoverflow.com/questions/38640109/logistic-regression-python-solvers-defintions
2. https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8
3. https://medium.com/analytics-vidhya/adjusting-for-inflation-when-analysing-historical-data-with-python-9d69a8dcbc27
4. https://github.com/YashMotwani/TMDB-Movies-Dataset-Investigation-/blob/master/TMDB_Movies_Dataset_Analysis.ipynb
5. https://medium.com/analytics-vidhya/implementing-linear-regression-using-sklearn-76264a3c073c
6. https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8
7. https://github.com/jeremy-lee93/dsc-mod-1-project-v2-1-onl01-dtsc-pt-052620
8. https://www.justintodata.com/logistic-regression-example-in-python/
9. https://stats.stackexchange.com/questions/463690/multiple-regression-with-mixed-continuous-categorical-variables-dummy-coding-s
10. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.plot_precision_recall_curve.html
11. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
12. https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/

THANK YOU

Do you have any questions?
Please checkout our notebook too, there's additional questions there.

Ameeshi Gupta (AMEESHI001@e.ntu.edu.sg)
Darren Chan Inn Siew (DCHAN025@e.ntu.edu.sg)
Koh Jia Cheng (C200172@e.ntu.edu.sg)