TV Viewership Age Demographics

A CASE STUDY

Contents

EMPATHIZE

In this phase we will try to understand the problem, the challenge that we are trying to solve and how we are going to address it.

IDEATE

In this phase we will try to build the features and explore and analyse the data.

PROTOTYPE

In this phase we will be building training, testing and evaluating our models to measure its performance

IMPLEMENT

In this phase we will propose what to do with the models.

Disclaimer: I have tried to approach the whole presentation in a non traditional format, using principles of Design thinking to add a story telling framework as well as an exploratory approach

Modern Day Cable TV Services:

TV viewership has evolved over time from traditional analog cable television services to present day, DTH based digitally integrated services, offering a multitude of options and features at the fingertips of the viewer.

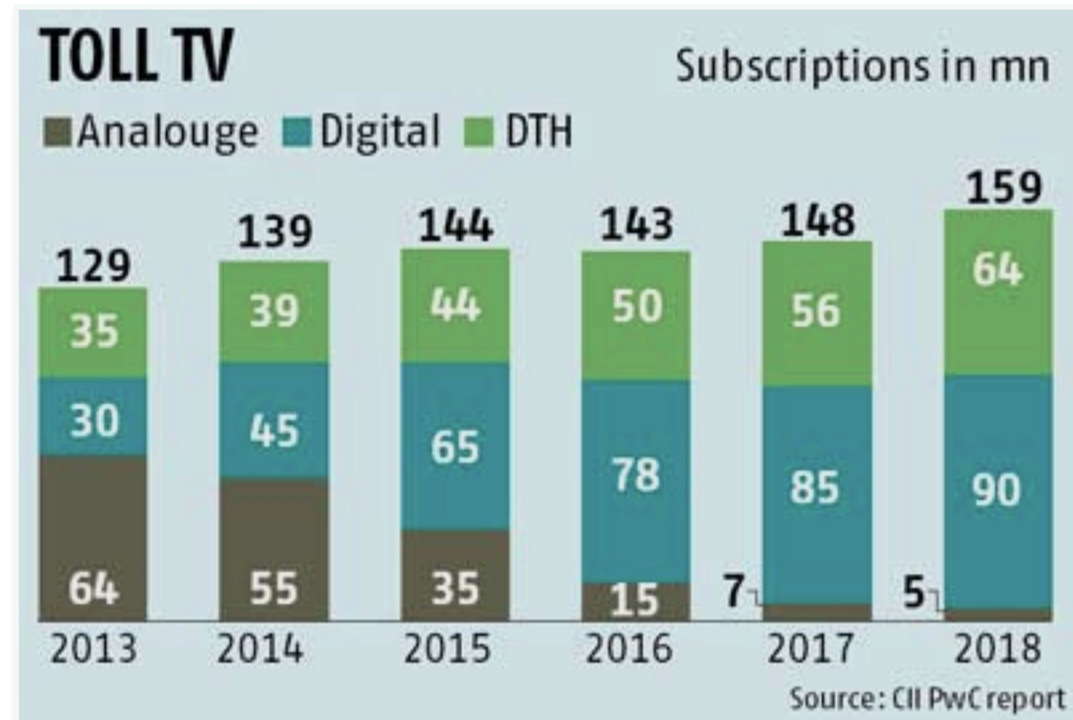
Modern Day DTH Services

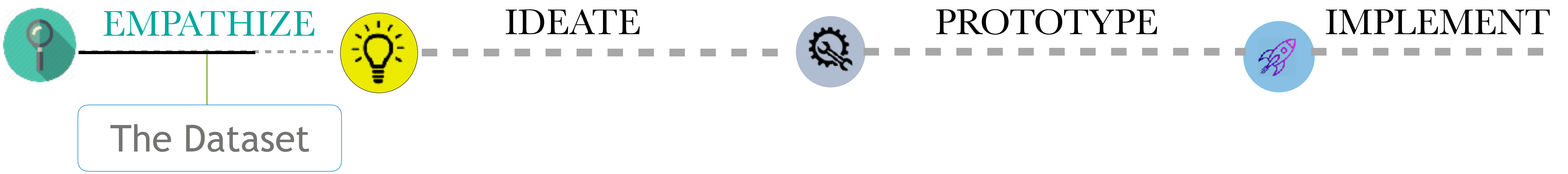
EMPATHIZE IDEATE PROTOTYPE IMPLEMENT

Context

What is the context of doing a tv viewership demographics ?

In the digital era, companies are moving towards AI driven technologies where they can understand their customers, segment them on basis of their viewership patterns, and provided recommended content based on these qualitative behaviour of their users.
In the best interest of the business, providing correct content to reduce churn is the main motive behind building an intelligent age classification system.





What does our Dataset speak ?

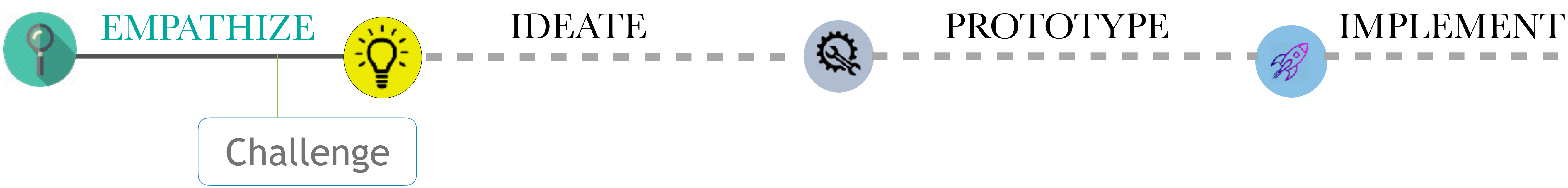
household id	session start	session end	channel name	title	original broadcast start	original broadcast end	session type	session sub type	genre	sub genre	playback speed	episode title	series title	gender	dob
432215006	2016-04-09 08:36:52	2016-04-09 08:51:52	Nick Junior	Paw Patrol	2016-04-06 10:30:00	2016-04-06 10:45:00	TIMESHIFT	SERIES LINK BOOKING	Kids/Youth	For ages 6-14	1000	Pups and the Mischievous Kittens		Please Spec	
432215006	2016-04-27 10:03:48	2016-04-27 10:08:48	Nick Junior	Paw Patrol	2016-04-27 10:00:00	2016-04-27 10:15:00	TIMESHIFT	BUFFER	Kids/Youth	For ages 6-14	1000	Pups Save a Sniffle		Please Specify	1900-01-01
432215006	2016-04-27 17:03:48	2016-04-27 17:03:48	Nick Junior	Paw Patrol	2016-04-27 17:00:00	2016-04-27 17:15:00	TIMESHIFT	BUFFER	Kids/Youth	For ages 6-14	1000	Pups Save a Sniffle		Please Specify	1900-01-01
432215006	2016-04-18 09:02:46	2016-04-18 09:17:46	Nick Junior	Paw Patrol	2016-04-18 08:45:00	2016-04-18 09:00:00	TIMESHIFT	BUFFER	Kids/Youth	For ages 6-14	1000			Please Specify	1900-01-01
432215006	2016-04-22 17:28:45	2016-04-22 17:28:45	Nick Junior	Paw Patrol	2016-04-22 17:15:00	2016-04-22 17:30:00	TIMESHIFT	BUFFER	Kids/Youth	For ages 6-14	1000			Please Specify	1900-01-01

We are provided a dataset in the CSV format that contains the following columns.

['household_id', 'session_start', 'session_end', 'channel_name', 'title','original_broadcast_start', 'original_broadcast_end', 'session_type','session_sub_type', 'genre', 'sub_genre', 'playback_speed','episode_title', 'series_title', 'gender', 'dob']

These columns help us to answer the following context:

- What is the household id of the user?
- When did they start a session ?
- What tv channel did they watch ?
- What did they view in the channel ?
- When did they have their signals broadcasted ?
- What type of session was the broadcast ?
- What type of session sub type was it ?
- What kind of genre did it fall under ?
- What kind of sub genre did it fall under ?
- What was the title of the episode ?
- What was the title of the series ?
- What is the gender of the viewer ?
- What is the dob of the client ?



The Challenge:

So now that in the previous steps of this phase we have understood the following information:

- What is the business context?
- Who are my stakeholders?
- What information is in the dataset?

The challenge therefore is:

Can we predict the user age/age group based on their viewership behaviour ?

```
*** Some of the information gained from the dataset **

How many household ids are there ? 3863142
How many household ids are unique ? 2813
How many channels are there ? 238
Number of unique titles in total ? 12627
Types of sessions available? {'LIVE', 'REPLAY', 'VOD',
'TIMESHIFT'}
Types of sub sessions available? {'LIVE', 'REPLAY', 'TVOD',
'SVOD', 'SINGLE RECORDING', 'BUFFER', 'MYPRIME', 'SINGLE TIME
BASED BOOKING', 'REPEAT TIME BASED BOOKING', 'SERIES LINK
BOOKING'}
How many genres are there? 14
How many sub-genres are there? 72
How many types of playback speed are there? 793
How many episode titles are there? 22628
How many series title are there? 5056
How many gender types are there? {'Please Specify', 'To be
specified', 'Female', 'Male'}
How many users have unique dob? 987
```

EMPATHIZE IDEATE PROTOTYPE IMPLEMENT

Tackling the Outliers: Preprocessing Step I.

Detecting and correcting outliers:

Our dataset is extremely huge (more than 3.8 million rows with 16 different features). Handling outliers is the first task in our preprocessing steps. On close inspection, we get to see that the following information:

1. The NaN rows

(Fig 1: We can see that columns ‘Episode_title’, ‘Series Title’, ‘Original Broadcast Start’, ‘Original Broadcast end’, ‘Playback Speed’ and ‘Title’ has empty values present with 2.8million rows in ‘Episode_title’ being empty rows. This accounts for more than till 73% of the total rows in the dataset)

2. The non imputable rows :

(Fig 2: 65% of the DOB are entered as 1900. It is extremely hard to understand a mean age and apply statistical imputations to replace this, as it will tip the scale towards this bias.)

Action taken: Dropped rows that are corrupted, cannot be replaced for a clean dataset.

	index	0
0	household_id	0
1	session_start	0
2	session_end	0
3	channel_name	0
4	title	5
5	original_broadcast_start	12181
6	original_broadcast_end	12181
7	session_type	0
8	session_sub_type	0
9	genre	0
10	sub_genre	0
11	playback_speed	782
12	episode_title	2805111
13	series_title	1771372
14	gender	0
15	dob	0

Fig 1.

	index	dob
0	1900-01-01	2568362
1	1970-01-02	10216
2	1981-02-24	8458
3	1960-01-06	7020
4	1942-11-21	6528
5	1955-08-17	6381
6	1964-10-25	6349
7	1982-04-27	5974
8	1976-05-08	5634
9	1970-11-07	5496

Fig 2.

EMPATHIZE IDEATE PROTOTYPE IMPLEMENT

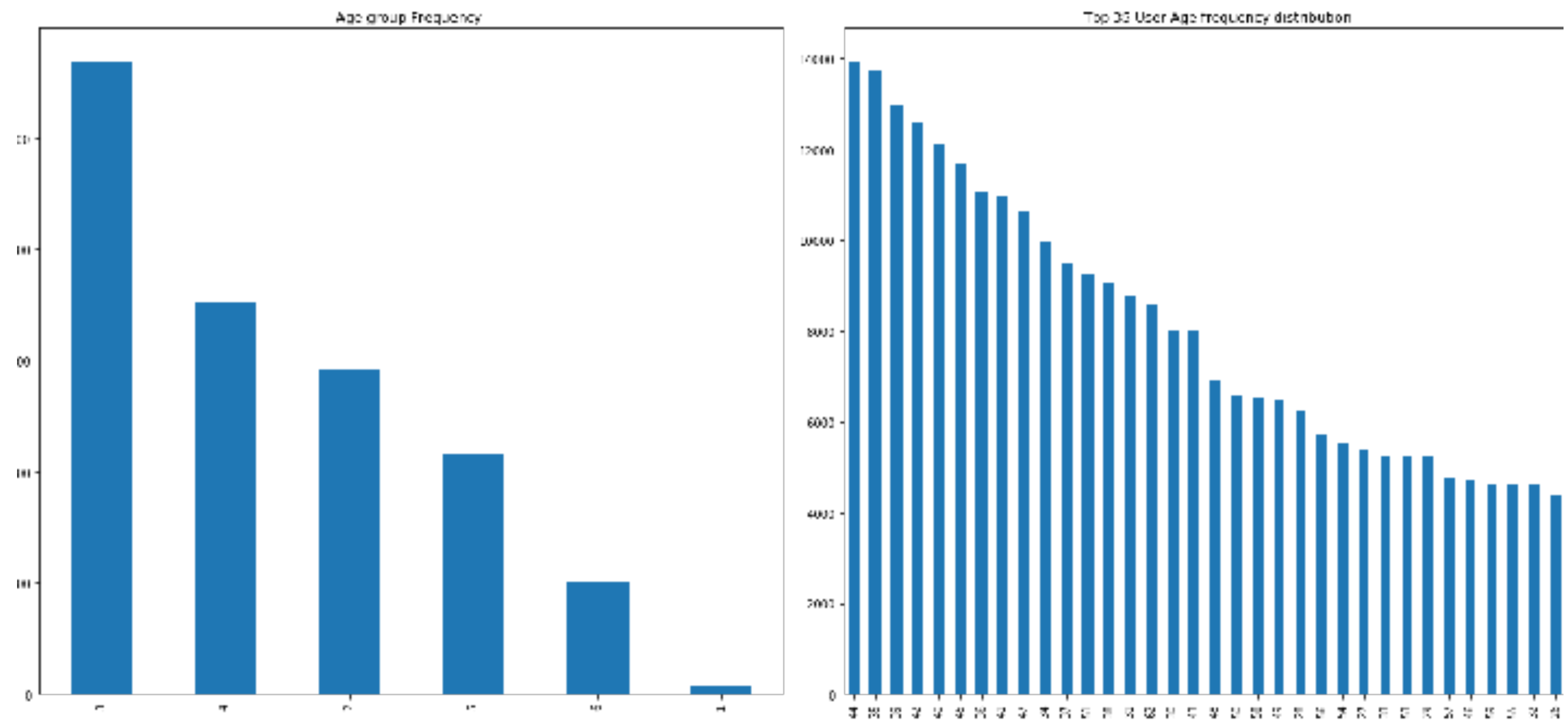
Information from the cleaned dataset I: Age Groups

The Age Groups:

Based on the new cleaned dataset, we have calculated the ages of the users from their DOB, and have aggregated them into 6 different age groups, listed as follows:

1. <25
2. 25-35
3. 35-45
4. 45-55
5. 55-65
6. 66+

The chart shows that the middle age, 35-45 age group is the modal class in the distribution.
Assuming that majority of the users will be having a family. We will try to establish a link in the following slide.



Ranked Age group and Age distribution of the users

EMPATHIZE IDEATE PROTOTYPE IMPLEMENT

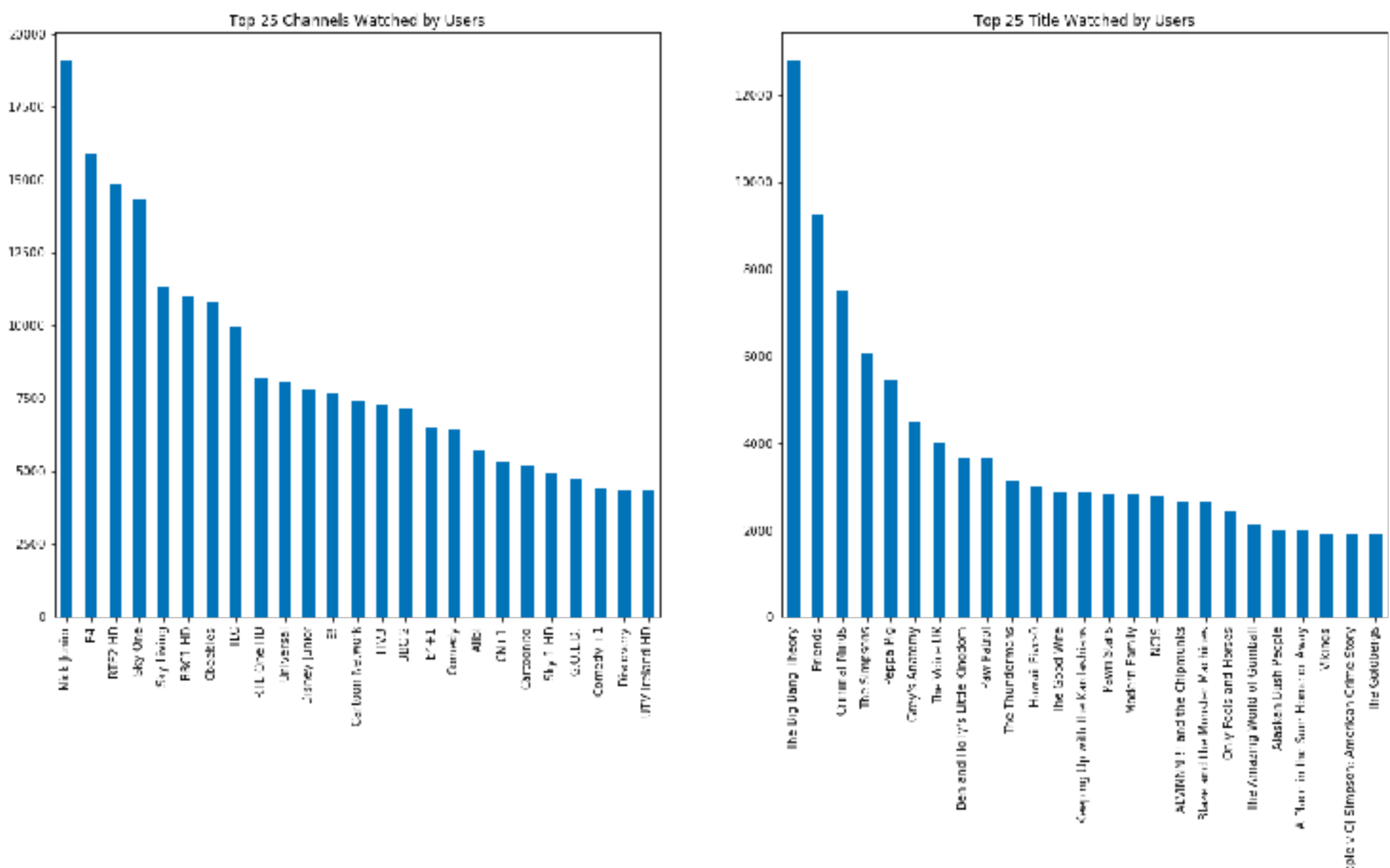
Information from the cleaned dataset II: Most Channels and Titles watched

Introspective into the viewing pattern:

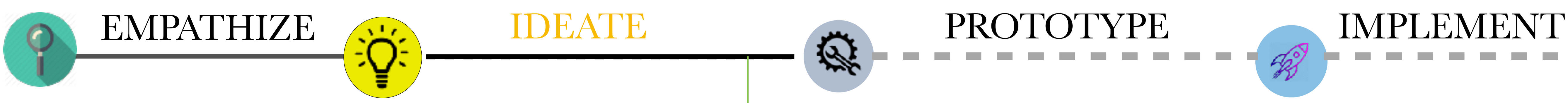
In the previous slide we had tried to establish that since the modal age group is middle age, we are assuming that this group of users have a family.

In this chart, from the channel we can see that most watched channel was Nick Junior, while the most watched title is the Big Band theory.

These kind of information is painting a picture of the qualitative viewing patterns of the users.



Top 25 Channels and Titles watched by Users



Preprocessing II: Encoding

Encoding:

In the last stage of this phase, we will try to encode our data.
The preprocessing and data transformation for our dataset is a bit tricky due to the following reasons:

Our dataset comprises of a lot of categorical columns like, ″Title″, ″Channel_name″, "Gender", ″Session_type″, ″Genre″, etc. Computers are not trained to process strings/ textual data and this information needs to be converted into numbers that can be used by a computer to build a model.
Essentially, there are 2 methods by which text can be converted to numbers:

- Word Vectorizer
- Label Encoding

Word Vectorizer converts words to word vectors using embeddings, but it is more applicable in a sequence of texts (example, text in a document). Since we don′t have text in a sequential format, this method is not advisable for our experiment.

Label Encoding on the other hand makes more sense as it maps and encodes each text to a value.

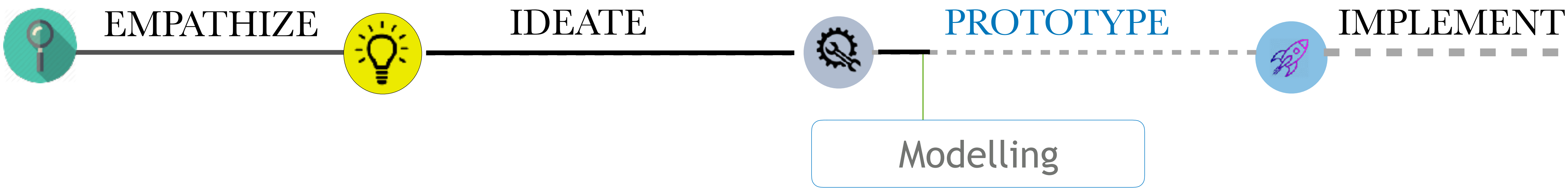
For example, if we are converting the titles, ″Titanic″, ″Avatar″, ″The Godfather″ to encoded labels, the computer will assign 1,2,3 for the titles respectively. The problem with this encoding is, the computer will assume that labels for 3 > 2 >1 (″The Godfather″ > ″Avatar″ > ″Titanic″) which is not the case. To solve this, we have to use OneHotEncoding. There is a drawback of increasing dimensionality while using One Hot Encoding technique, but for the moment we will not consider this, and use a limited number of rows to build and test our model.

* * One hot encoding transforms: a single variable with n observations and d distinct values, to d binary variables with n observations each. Each observation indicating the presence (1) or absence (0) of the dth binary variable.

```
df.dtypes
```

```
household_id         int64
channel_name        object
title               object
session_type        object
session_sub_type    object
genre               object
sub_genre           object
episode_title       object
series_title        object
gender              object
dob                 object
age_group            int64
user_age             int64
dtype: object
```

Non numeric columns having categorical information needs to be encoded



Modelling:

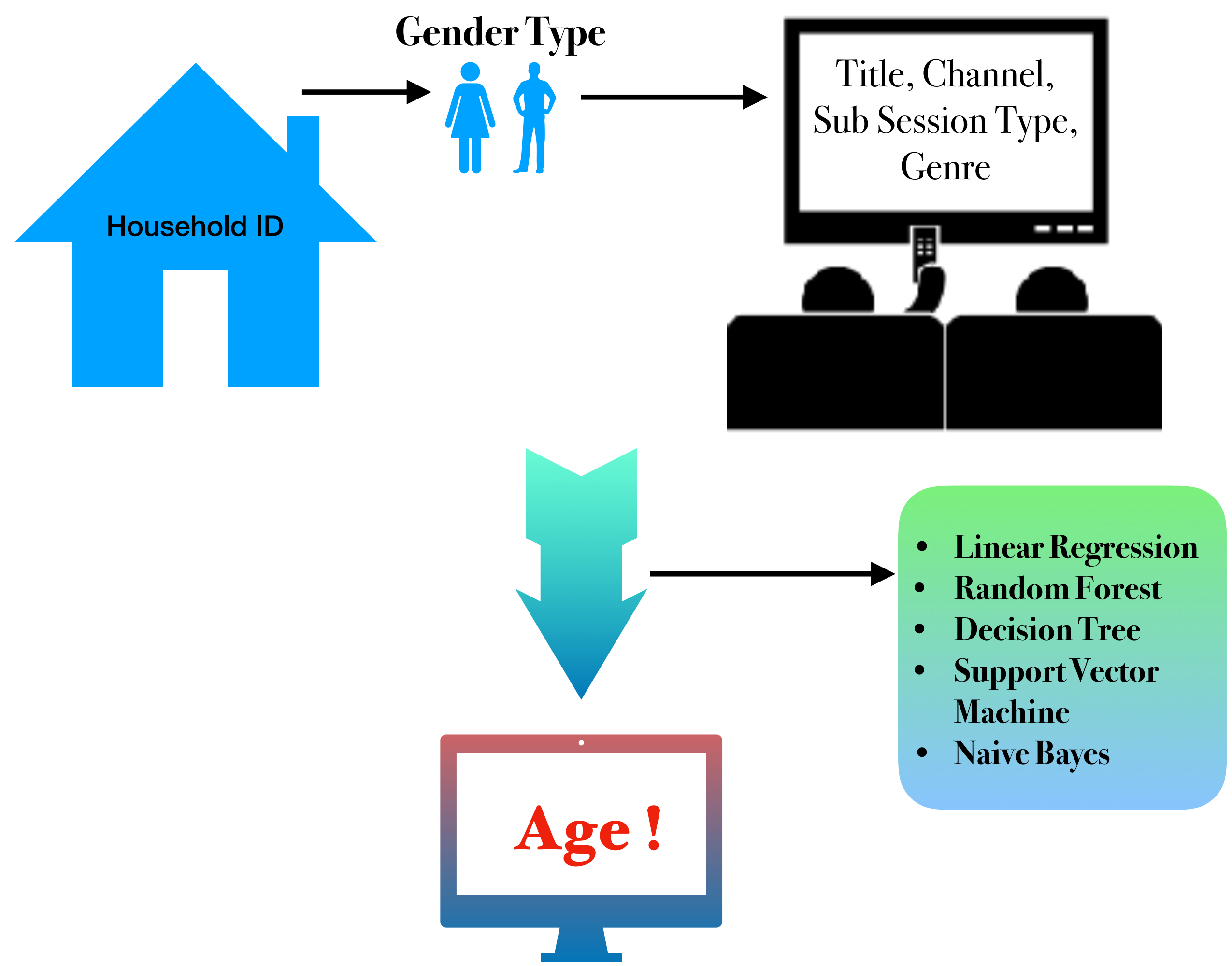
This is the stage where we will build a model on our encoded dataset. For the purpose of our experiment, I will try to establish a relationship between titles viewed, type of sub type session, genre and channel name patterns linked to a the household_id, and gender to predict and classify their age.

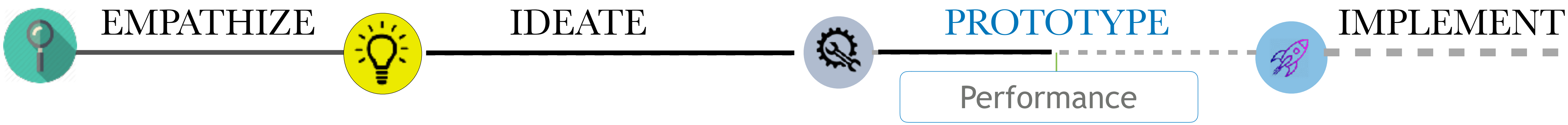
I have considered 2 approaches, for building this model.

1. Use logistic regression techniques to predict the age of user. Train the model on the above features with the response/ target variable being the 'User_age'. Try to measure if the model is able to predict the age of an user based on the features.

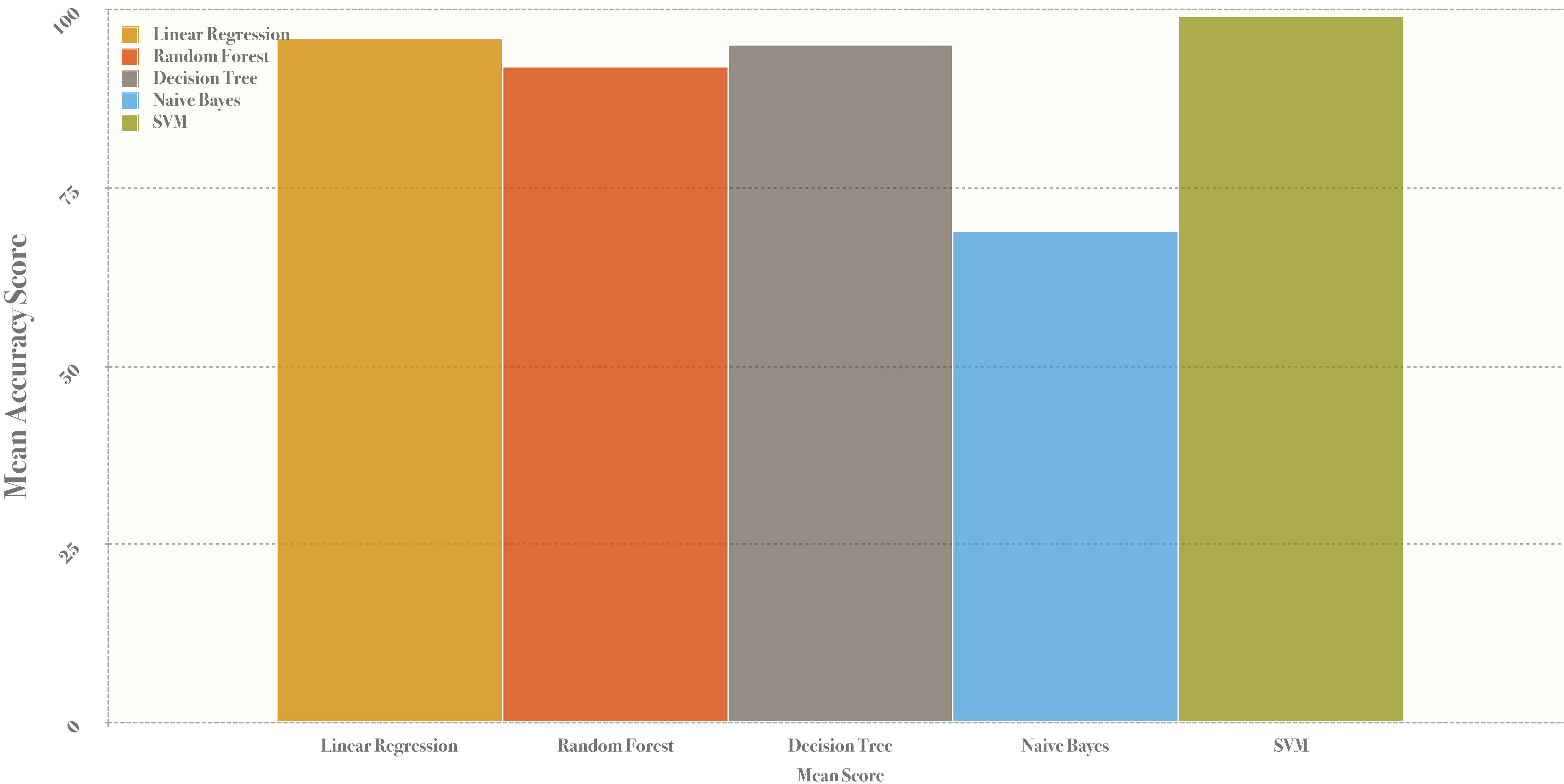
2. Use Classification techniques like Random Forest, Decision Tree, Support Vector Machine and Naive Bayes algorithms to train and test on the features and the target variable.

For this iteration, I am resorting to supervised machine learning techniques and for the sake of speed of computation, I have sampled 10K rows and will be training and testing on it.



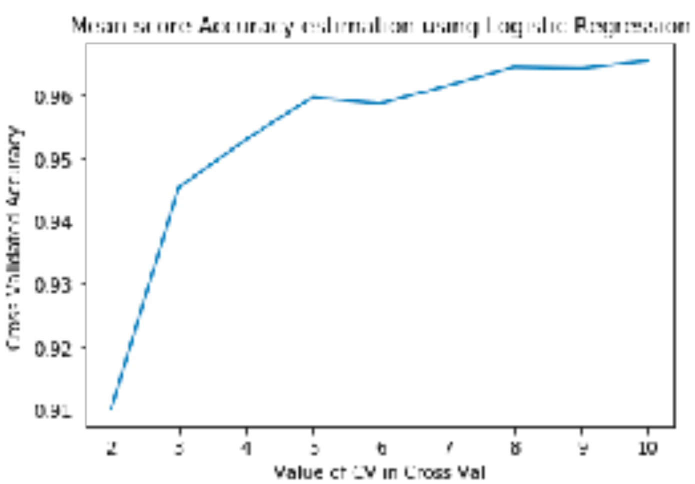


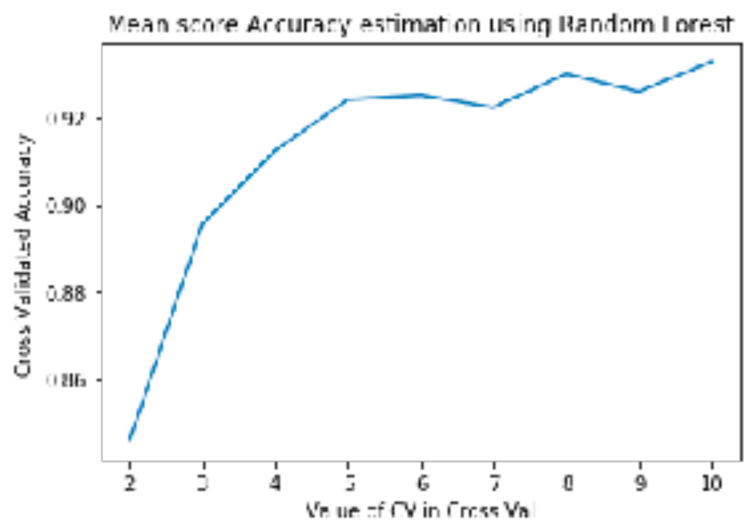
Mean Metric Scores of various models

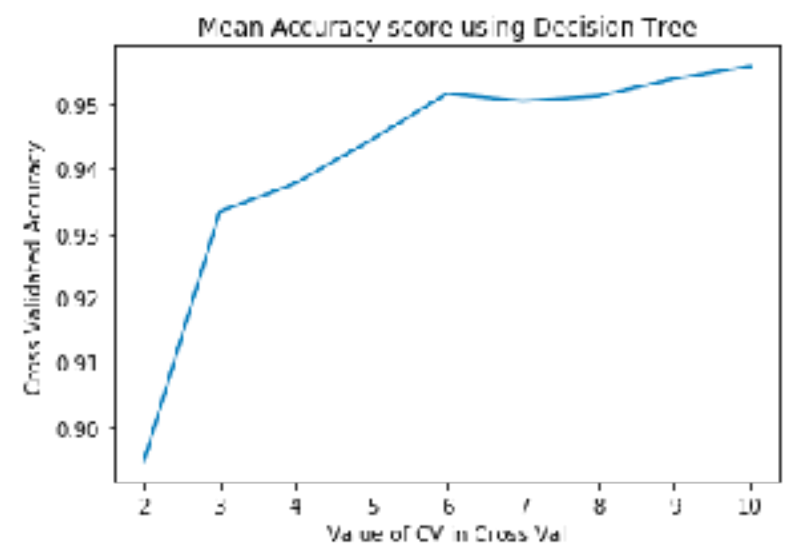


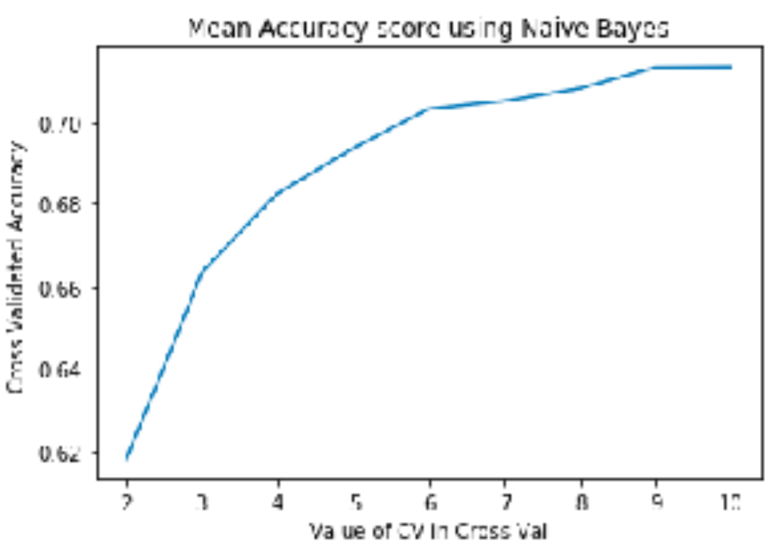
EMPATHIZE IDEATE PROTOTYPE IMPLEMENT

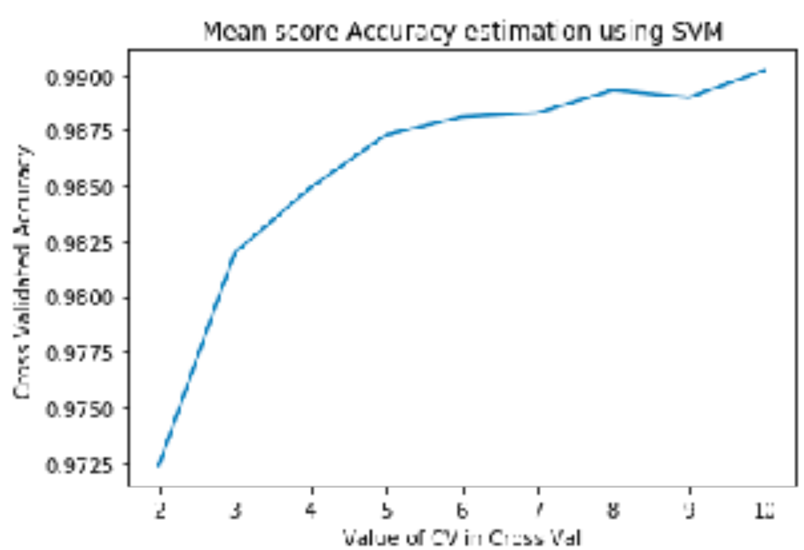
K Fold Cross Validation



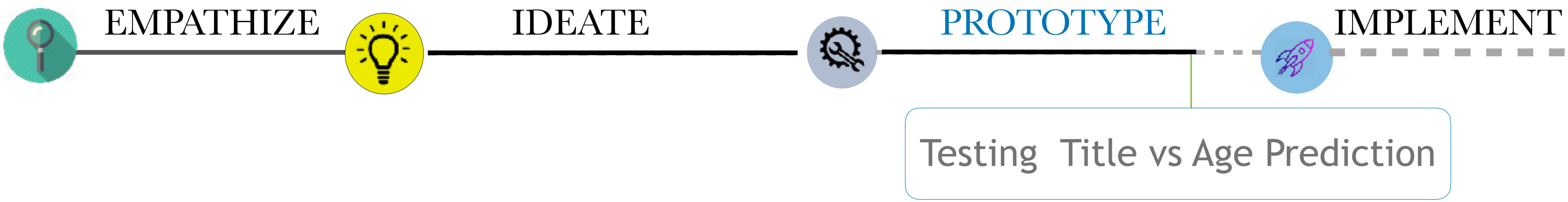


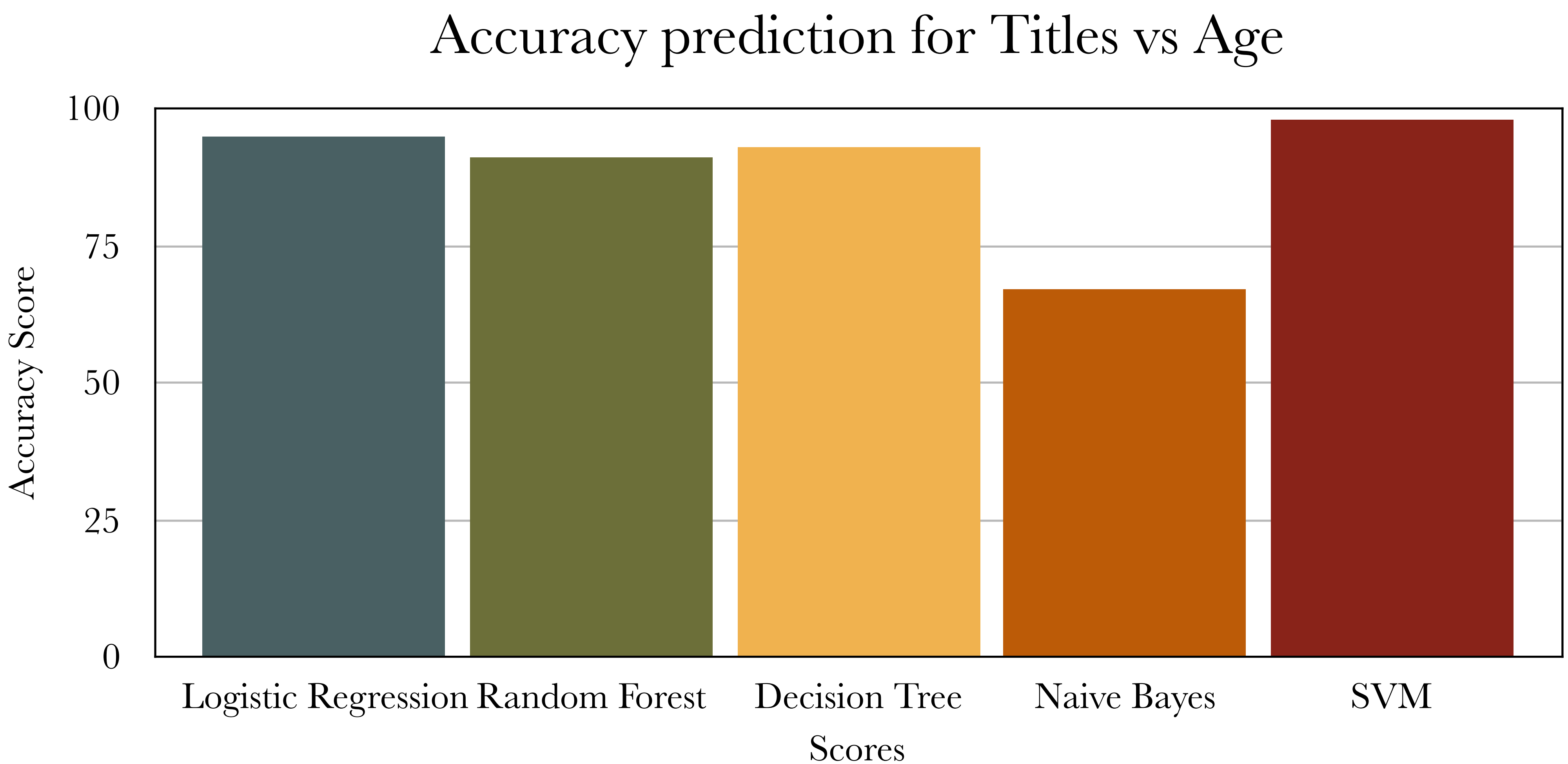




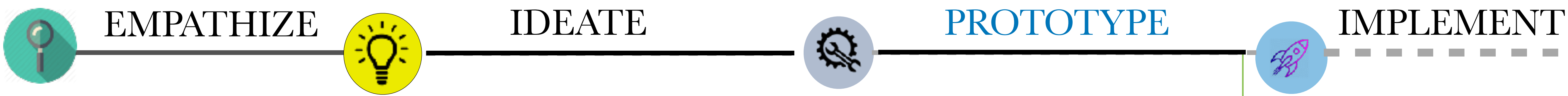


We are implementing K Cross Fold Validation to predict the mean accuracy of our models. Except Naive Bayes, the rest of the models have a very high accuracy score to predict the age of an user.





We can clearly see that our model (except Naive Bayes) has a very high accuracy to predict an age of a viewer based on the title that they have viewed



Building Title recommender

Title Recommender:

Here in this final step, I've tried to establish an item recommender based on titles watched by a household_id.

The assumption behind building this item similarity recommender are:

1. A household can have multiple viewers of multiple age groups.
2. If we can predict potential titles, it gives us an idea of the age group based on the content of the title.
3. We can use these predictions to predict the age and the age group.

```
--------------------------------------------------------------------
Training titles for the user userid: 432216770:
--------------------------------------------------------------------
LA Clippers Dance Squad
Hawaii Five-0
I Am Cait
Empire
My 600lb Life
Charmed
Friends
Angie Tribeca
The Inbetweeners
South Park
How I Met Your Mother
Bob's Burgers
The Hairy Bikers' Pubs That Built Britain
Inside Obama's White House
Royal Pains
Grey's Anatomy
Brooklyn Nine-Nine
Mike & Molly
Modern Family
Moone Boy
The Middle
The Next Great Baker
The Big Bang Theory
Frasier
Bad Education
A Place in the Sun: Winter Sun
Lip Sync Battle
Rude Tube
The Green Green Grass
Fat Chance
Escape to the Country
A Place in the Sun: Home or Away
The Goldbergs
Family Guy
Say Yes to the Dress
LA Ink
Father Ted
Extreme Couponing All-Stars
Once Upon A Time
```

Training Titles on UserId 432216770

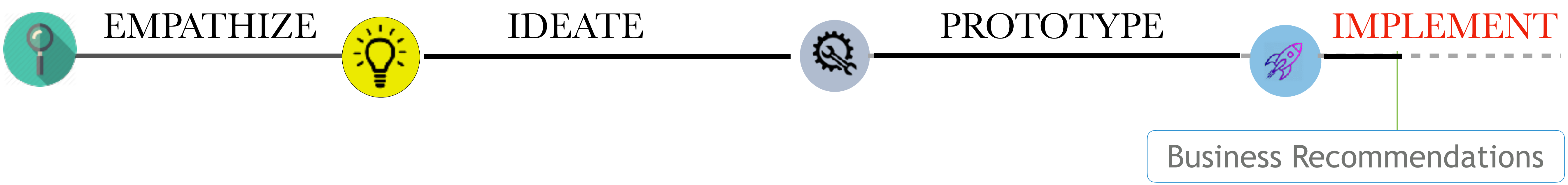
```
--------------------------------------------------------------------
Top 5 viewed items for household 432216770
--------------------------------------------------------------------

Friends                  74
Modern Family            14
The Goldbergs             5
The Big Bang Theory       5
Royal Pains               4
Name: title, dtype: int64
```

```
print("--------------------------------------------------------------------")
print("Recommendation process going on for household 432216770:")
print("--------------------------------------------------------------------")
is_model.recommend(432216770)
```

```
--------------------------------------------------------------------
Recommendation process going on for household 432216770:
--------------------------------------------------------------------
No. of unique titles for the user: 39
no. of unique titles in the training set: 1432
Non zero values in cooccurence_matrix :42942
```

	household_id	title	score	rank
0	432216770	The Simpsons	0.215203	1
1	432216770	Two and a Half Men	0.195386	2
2	432216770	Keeping Up with the Kardashians	0.194582	3
3	432216770	Criminal Minds	0.194532	4
4	432216770	Baby Daddy	0.191441	5
5	432216770	Only Fools and Horses	0.185190	6
6	432216770	Malcolm in the Middle	0.183604	7
7	432216770	Say Yes to the Dress: Atlanta	0.182253	8
8	432216770	Futurama	0.175461	9
9	432216770	CSI: Crime Scene Investigation	0.174231	10

Recommending Titles to userID 432216770 based on Item Similarity



Business Recommendations and Use Cases:

1. The title recommender can be used along with the models to predict suitable titles for varied age groups.
2. The potential to understand the demographic and viewing behaviour of the viewer can be used to target suitable ads.
3. Title based filtering can be used to filter ads for the targeted age groups/ gender.
4. Genre/Sub-genre based content recommendations can be improved

EMPATHIZE IDEATE PROTOTYPE IMPLEMENT

Conclusion and Future Scope for improvements

Conclusion:

To build a predictive model based on customer behaviour is not easy. A deep domain knowledge is required to understand the users.

Based on the assumptions that a viewer is not restricted to only person in the household, and trying to quantify his/her viewing qualitative viewing patterns. Our models can predict to 98% accuracy the age of the viewer.

In the second iteration based on the title vs age prediction, we are trying to establish a more 1:1 relationship between quantifying What the user watches to what their age can be.

Future Scope:

1. Use unsupervised methods to understand customer behaviour and gain more insights.
2. Use Ensemble approaches to improve the accuracy of the predictions.
3. Train and test with a larger dataset.
4. Use more advanced techniques for preprocessing.
5. Look into other alternatives for encoding, since One-Hot Encoding increases dimensionality of the dataset by a very high magnitude.
6. Other features by feature engineering can be explored to potential patterns and gain more insights about an user