Memrise User Analysis And Predictions

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

Introduction

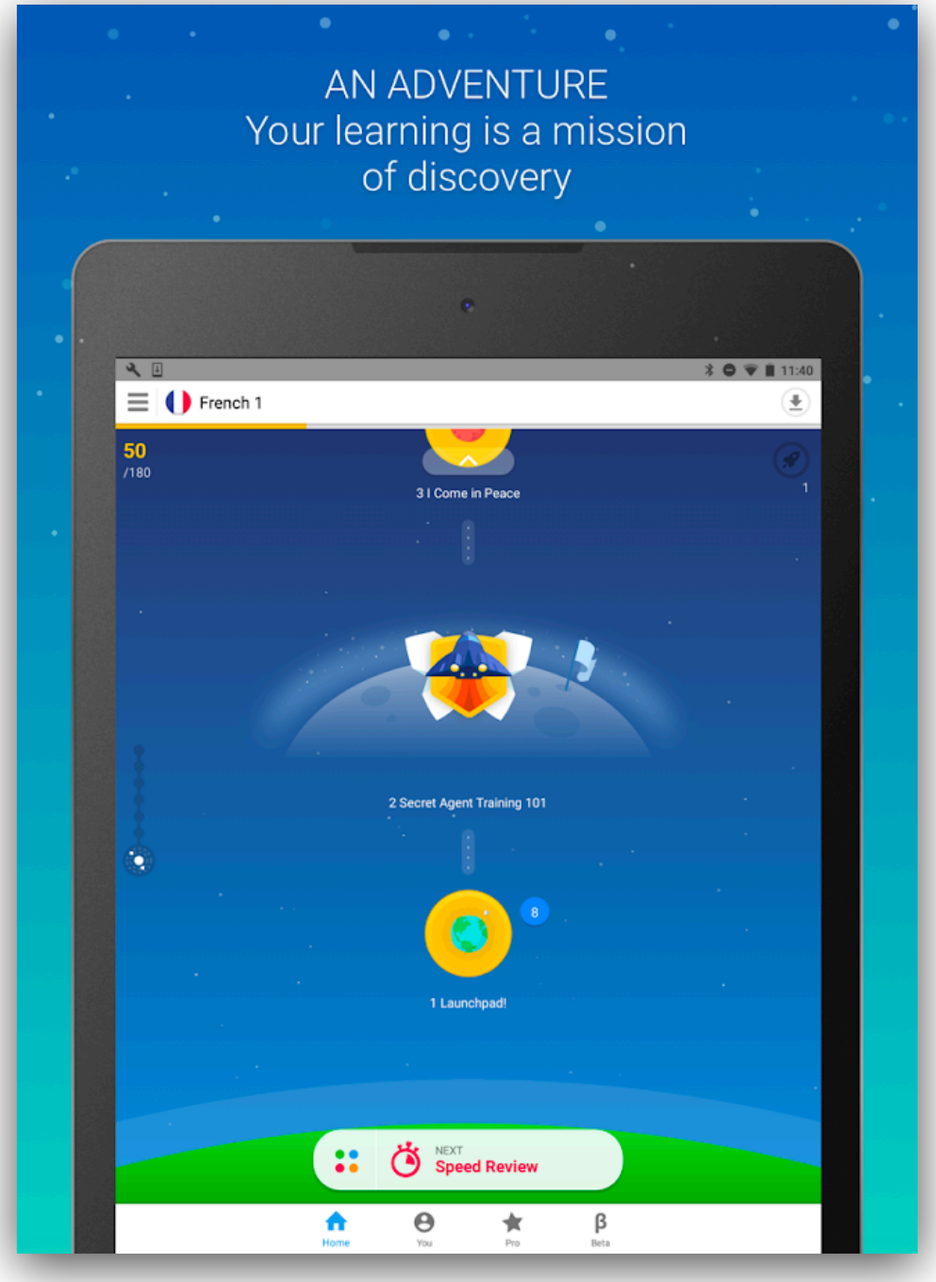
Lingua-franca:

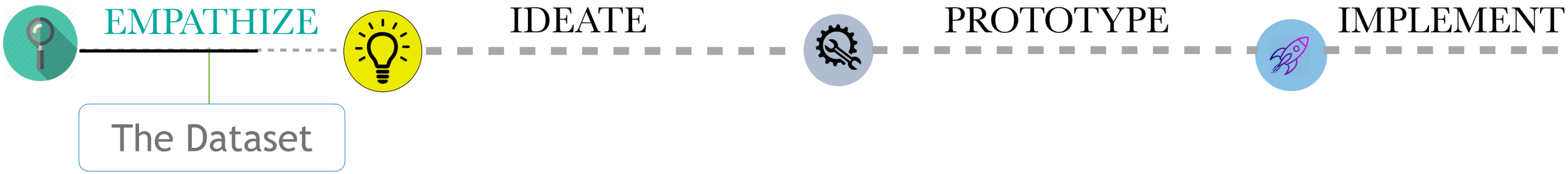
There are about 6909 different languages being spoken by the world's population. Language is the essence behind the progress of human race, it acts as a mode of communication, to express, to share, to understand. In short to, "Communicate".

Context

What is the context of this analysis ?

Empowering the varied lingo demographics to bridge each diaspora in this digital mobile age is an uphill task. To understand these groups, to assist them into learning a new language, and helping them overcome this challenge by analysing their behavioural traits is the premise behind this study.





What does our Dataset speak ?

We are provided 3 datasets in the CSV format that contains the following columns.

A. • **Users - Info about Memrise users gathered at time of signup**

○ user_id - randomised representation for a user

○ country - the country associated with the user's account

○ signup_app - the app used to sign up (memrise_[android/ios/web])

○ signup_time - the time when the user signed up

B. • **Learning sessions - A string of tests, can be of different types**

○ user_id - randomised representation for a user

○ learning_session_id - unique ID for the learning session

○ learning_session_type - the type of learning session

○ app - same as signup_app above

○ num_tests - the number of questions the user answered in this session

○ total_score - the total score for the session (score for each test is 0 - incorrect, 1 - correct)

○ start_time

○ completion_time

○ duration_mins - total time in minutes, excluding periods of inactivity

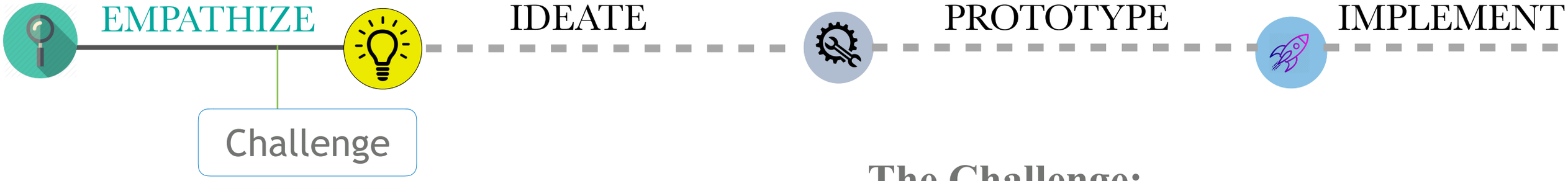
C. • **Subscriptions**

○ user_id - randomised representation for a user

○ period_month - the subscription period in months, ie 12 means yearly subscription

○ action - can be 'trial_started' or 'started' (user subscribes without trial)

○ action_time - the timestamp when the user subscribed



The Tasks therefore are:

- **How do the different data points relate to each other?**

- **Can you identify any groups of users based on their attributes? How do they differ?**

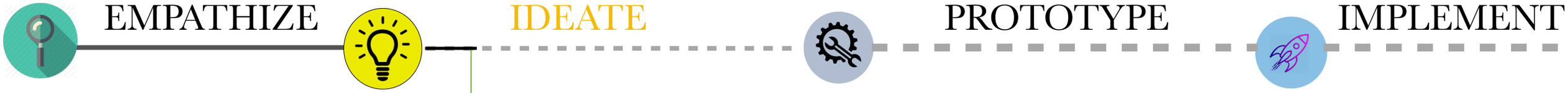
- **Do some users, or groups, seem to have a better learning performance than others?**

- **Are you able to predict the accuracy of a learning session? How confident are you with the prediction?**

The Challenge:

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

- Identify the stakeholders and provide recommendations
- Identify insights and how it connects to the business logic
- Possible implementation roadmap



Data Preprocessing:

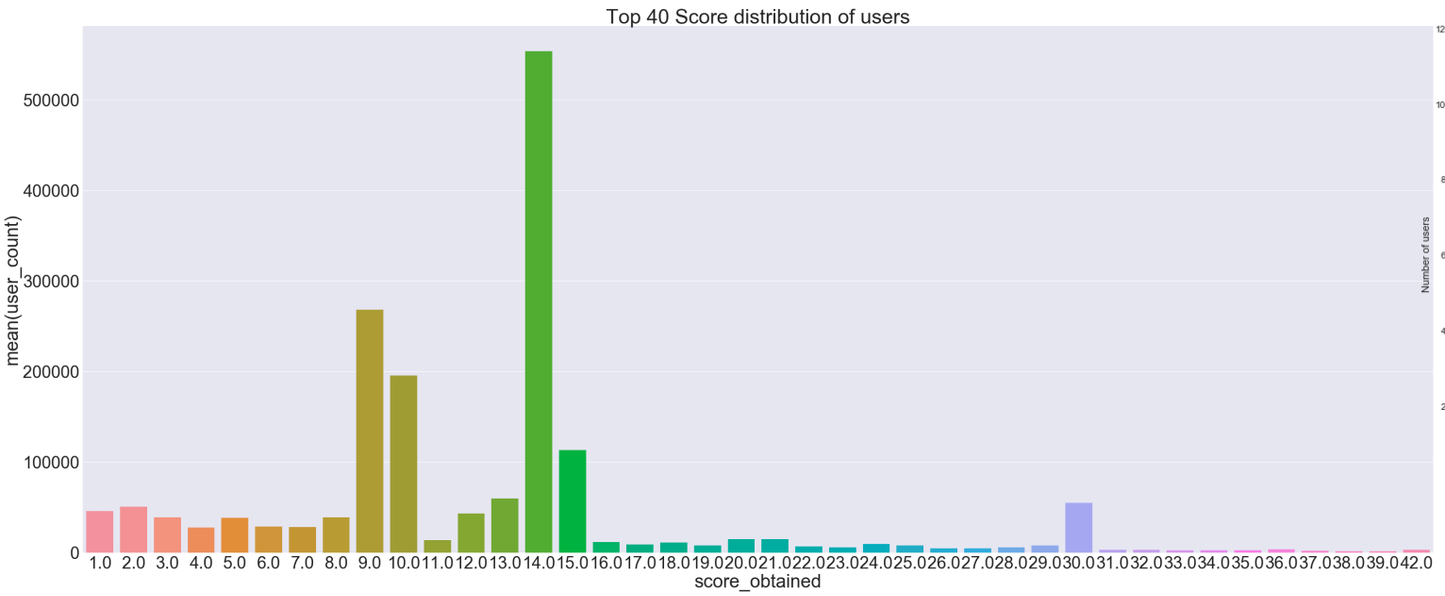
Our dataset is extremely huge (more than 1.8 million rows with 16 different features). Handling outliers is the first task in our preprocessing steps. On close inspection, we get to see that the following information:

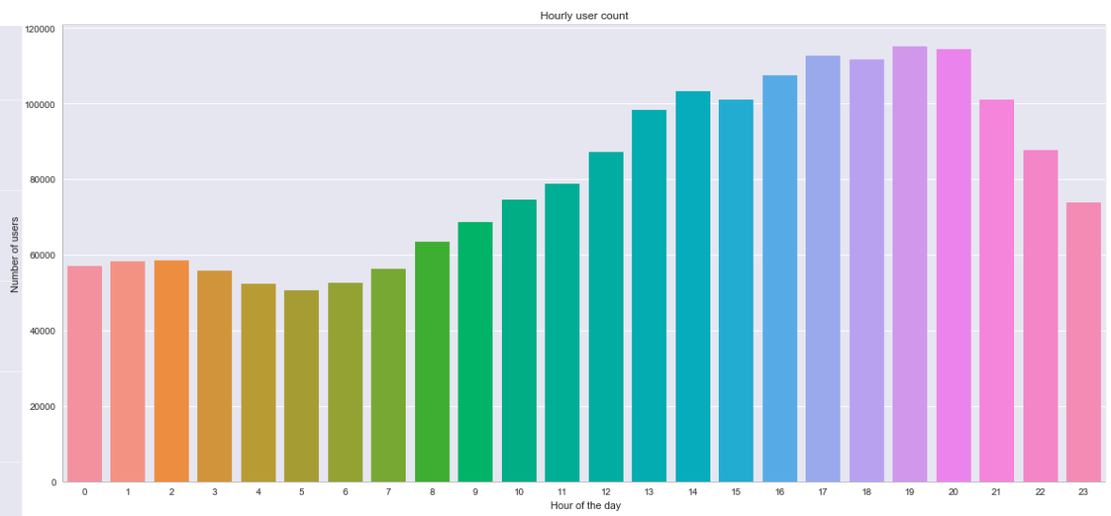
1. The NaN rows
   (Fig 1: We find around 17% of the learning dataset has null rows, and less than 4% for the other 2 datasets..

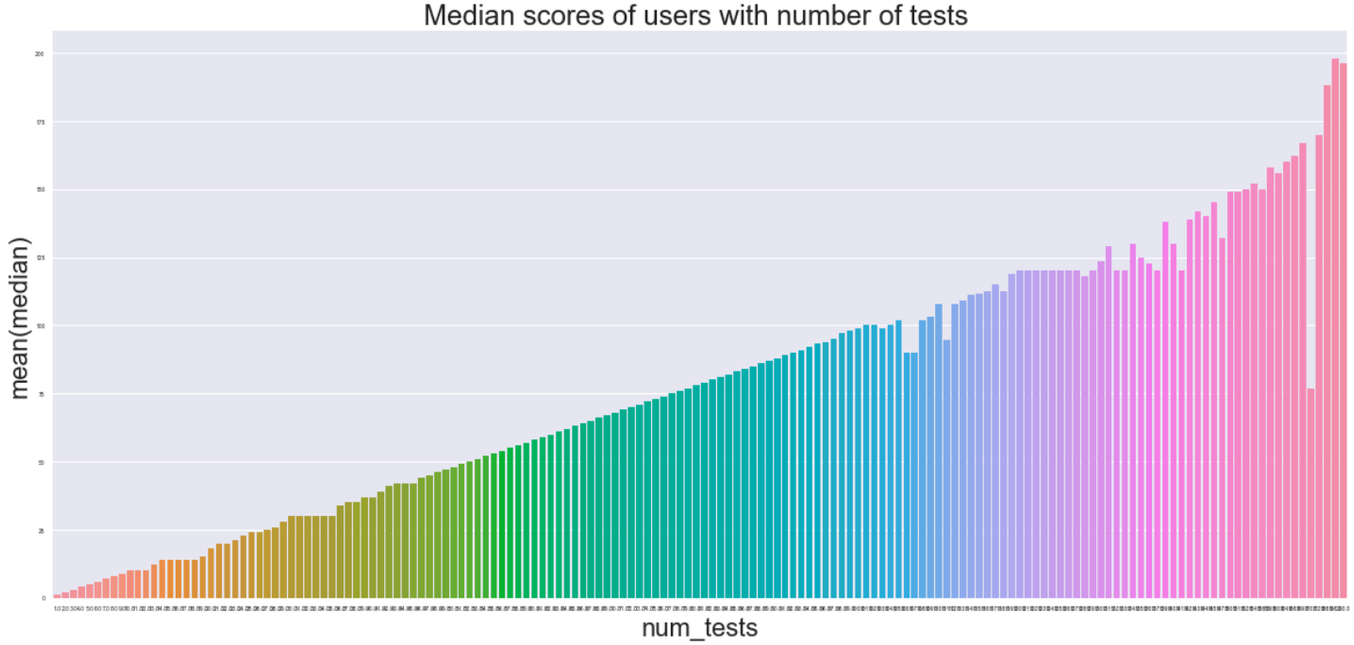
Action taken: Dropped rows that are corrupted, and cannot be replaced, for a clean dataset. Other alternatives could have been using, interpolation, and imputation techniques, but since the amount of corrupted set is fraction of the original dataset, I have chosen to drop the NaN rows to clean it.

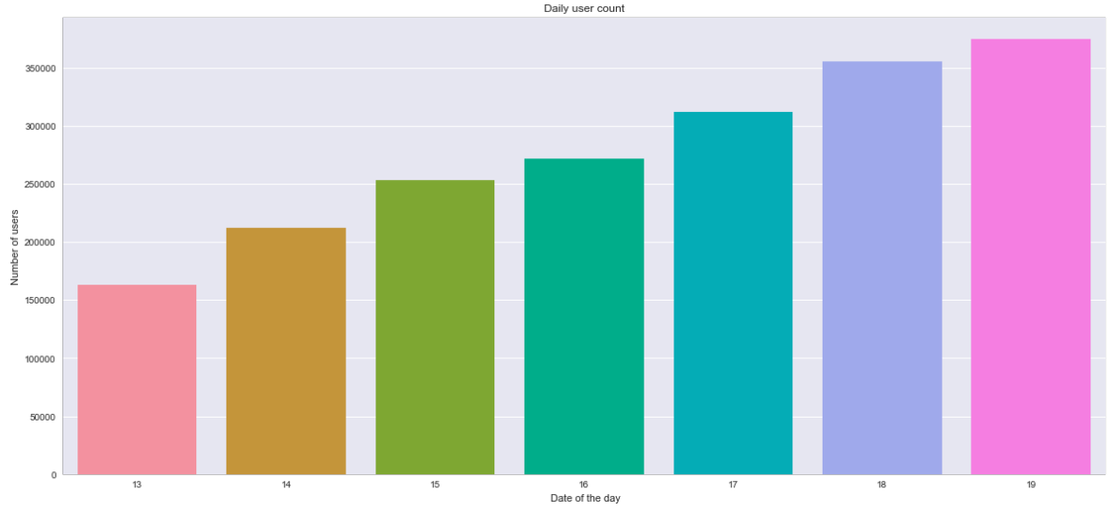
EMPATHIZE — IDEATE — PROTOTYPE — IMPLEMENT

Frequency Distribution Comparison



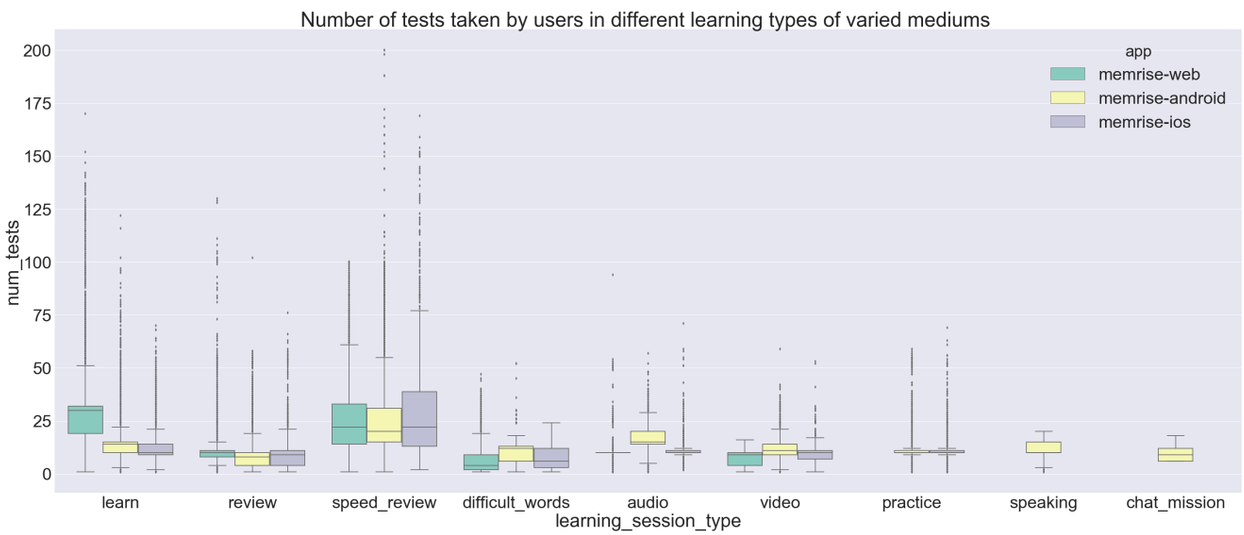


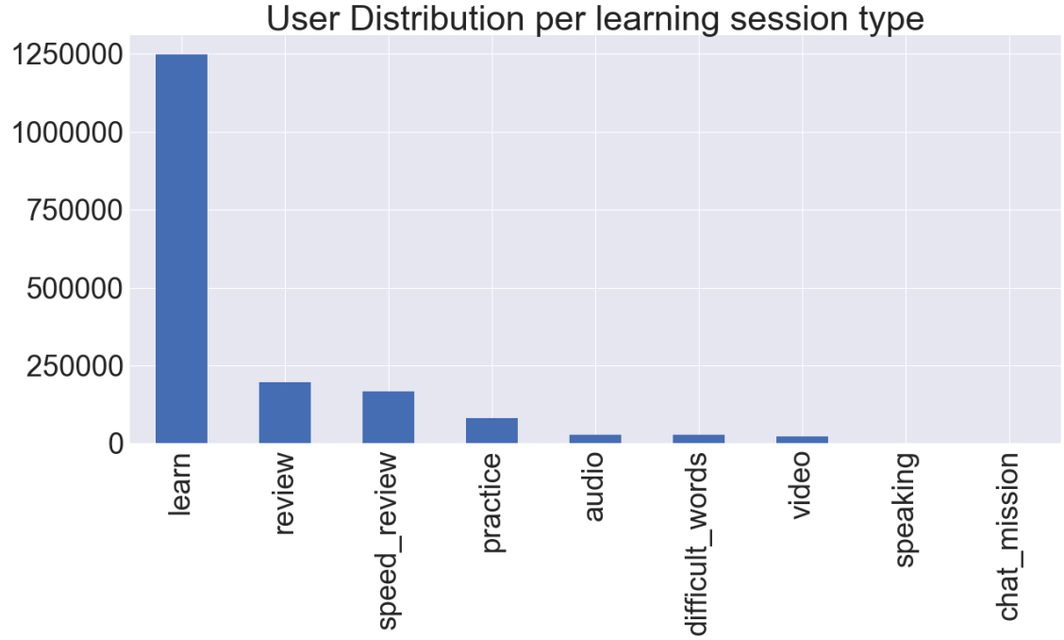


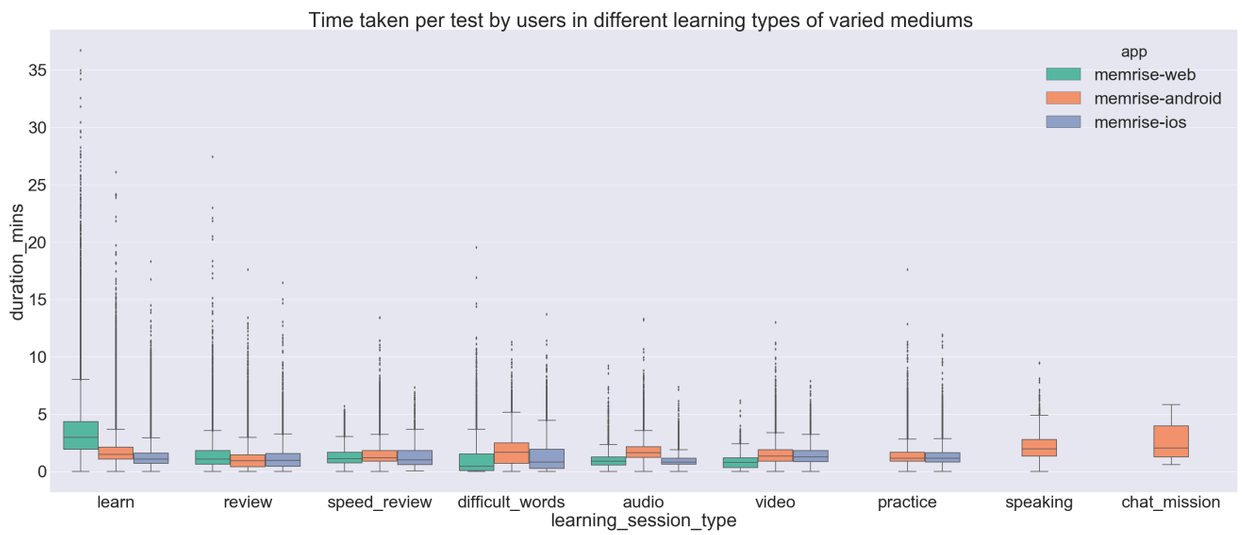


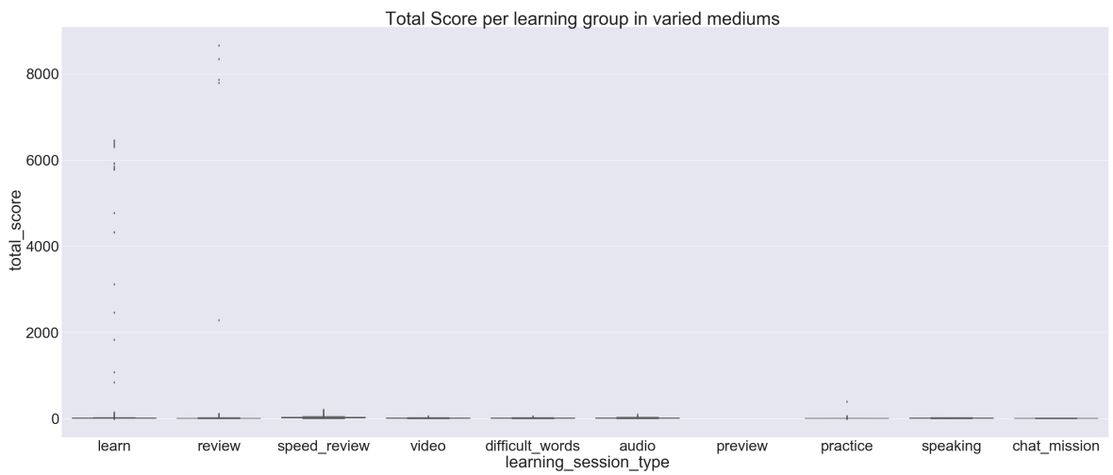
EMPATHIZE — IDEATE — PROTOTYPE — IMPLEMENT

Comparison of user performance stats in different learning phases



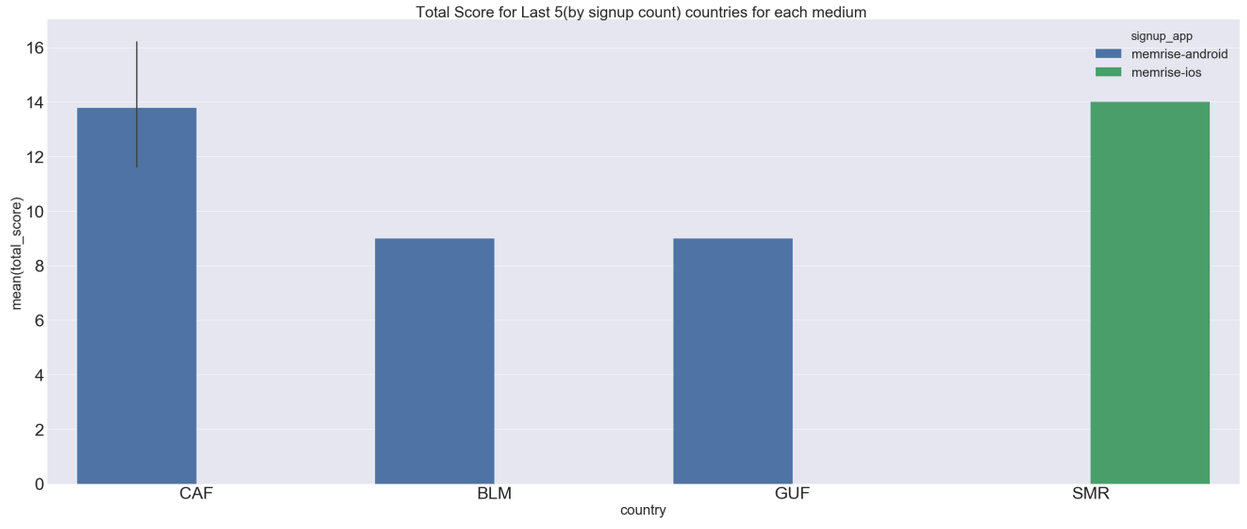


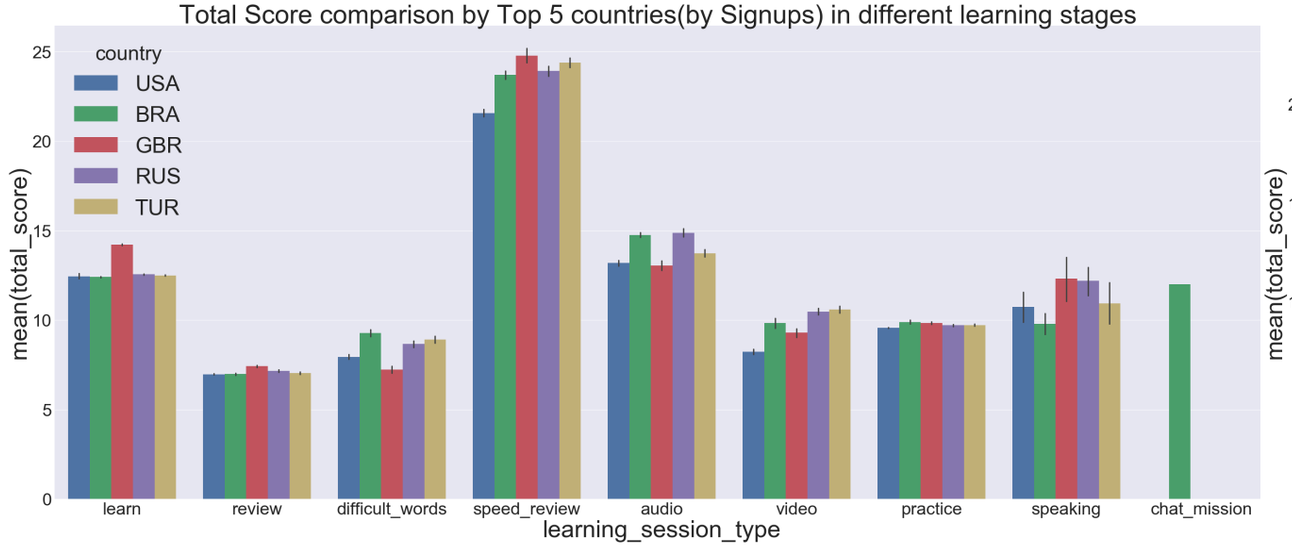


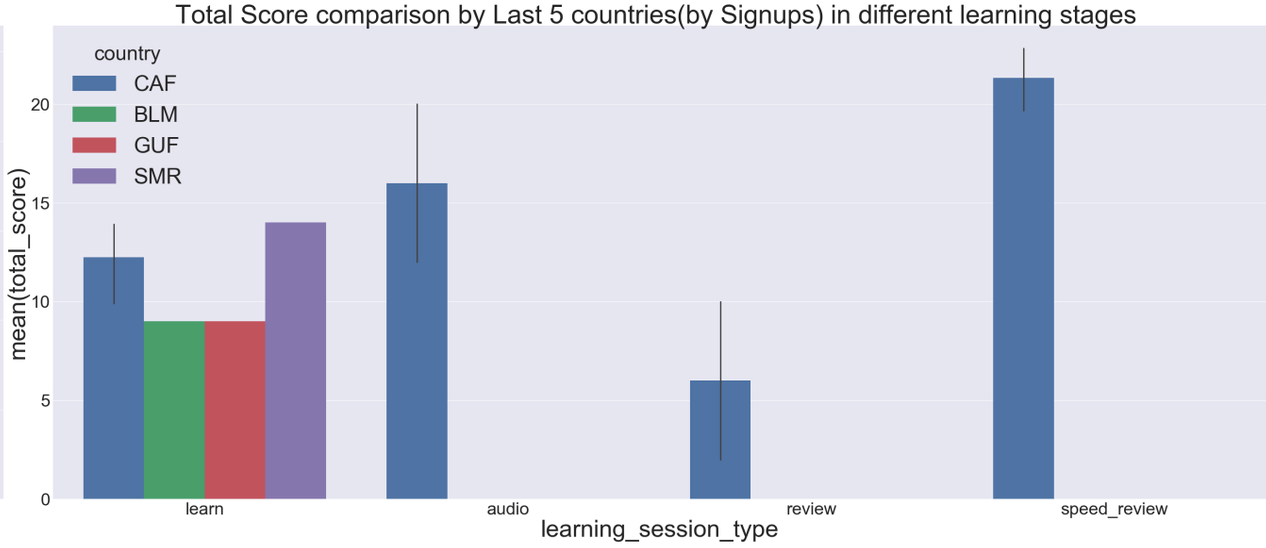


EMPATHIZE — IDEATE — PROTOTYPE — IMPLEMENT

Comparison of Top 5 and Last 5 countries

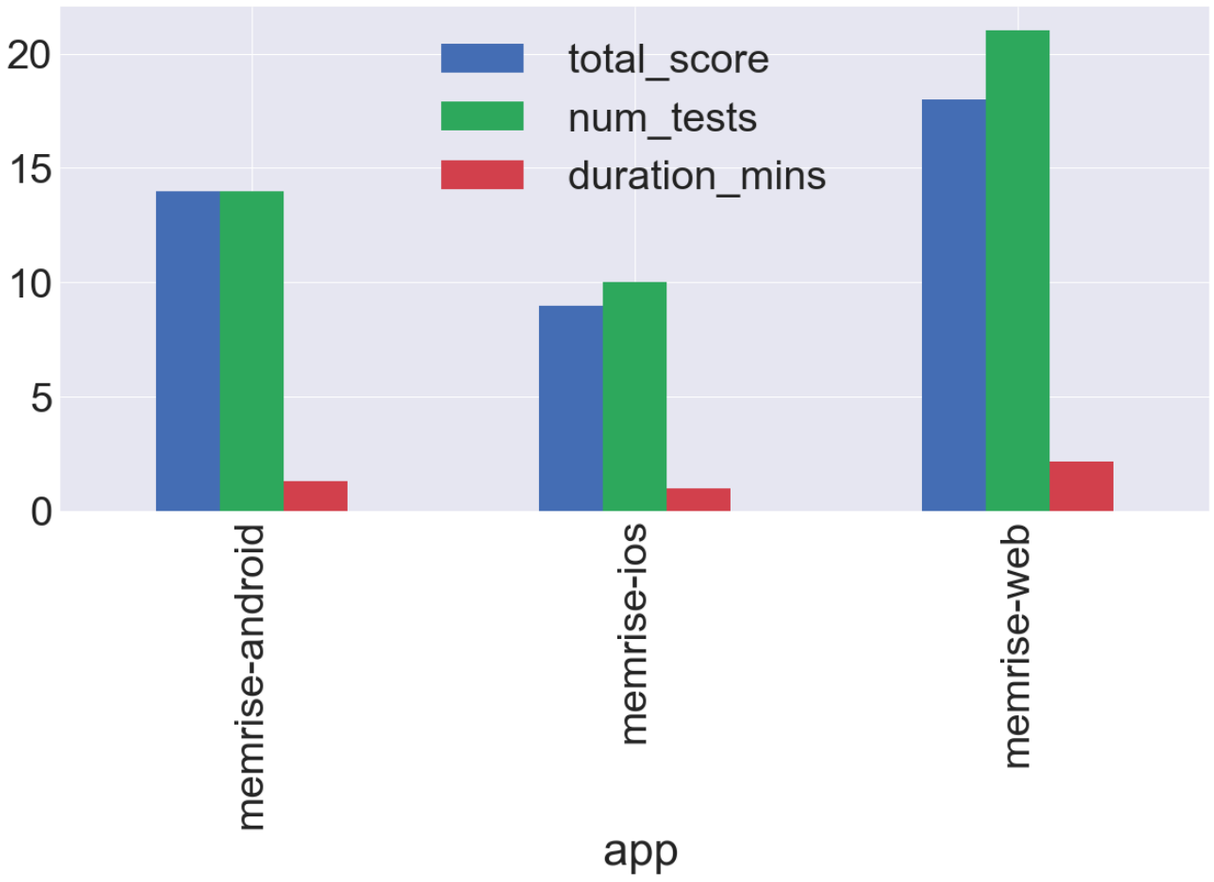


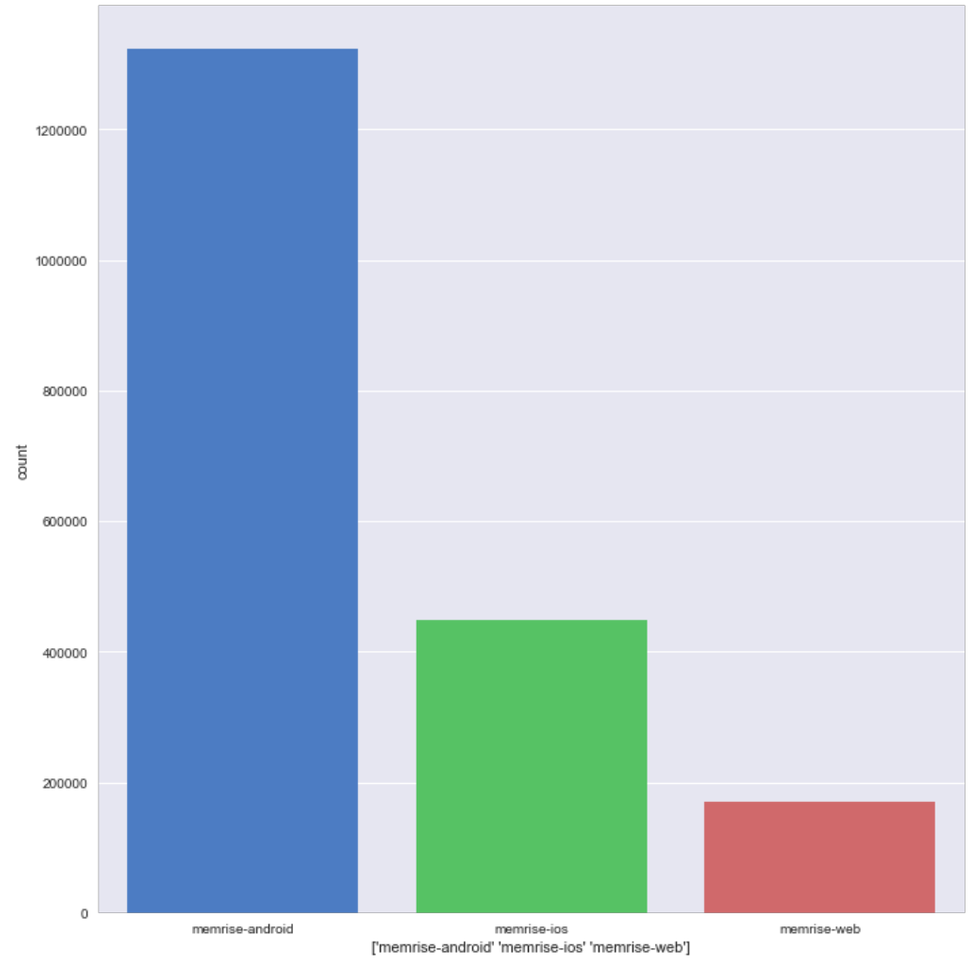


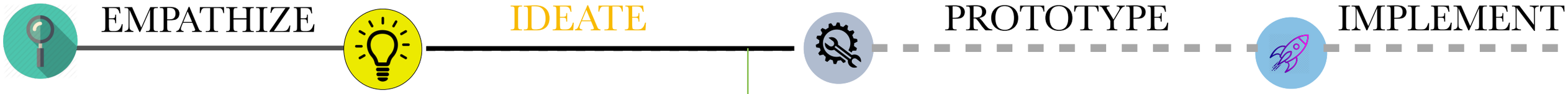


EMPATHIZE IDEATE PROTOTYPE IMPLEMENT

Comparison of groups in different medium channels:







Hypothesis Testing:

Is there statistical significance difference of user behaviour amongst different learning types?

1. Are the total scores of different learning types significantly different between the users? Yes
2. Are the number of test taken different between the learning types ? Yes
3. Is the time spent for the tests different between the different learning types? Yes

Conclusion: **There was significant difference in performance of users from different learning group types.**

Did the 2 top 5 and last 5 countries (by total signup count) users have statistical significance in their learning performance ?

1. Is the total score performance significantly different between the top 5 most countries with users and the last 5 countries with the least signed users ? No. Both of these groups performed the same.
2. Is the number of tests take significantly different between the top 5 most countries with users and the last 5 countries with the least signed users ? No. Both of these groups performed the same.
3. Is the duration spent for the tests significantly different between the top 5 most countries with users and the last 5 countries with the least signed users ? No. Both of these groups performed the same.

Conclusion: **There was significant difference in performance of users who had subscription duration of 1 months, 3 months and 12 months.**

Did the 2 different subscription users have statistical significance in their learning performance ?

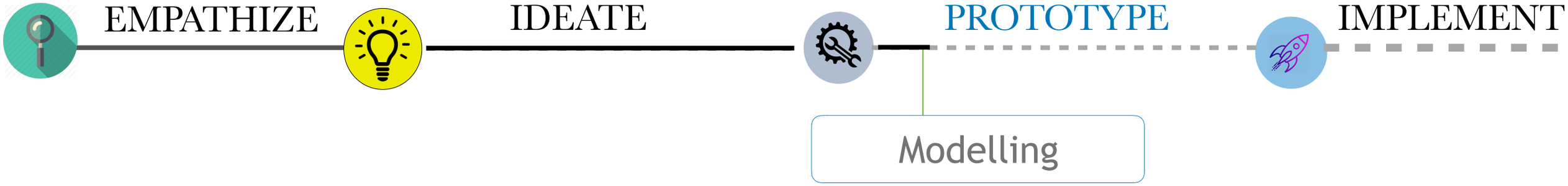
1. Is the total score performance significantly different between the 2 subscription groups? Yes.
2. Is the number of tests taken significantly different between the 2 subscription groups ? Yes
3. Is the duration spent significantly different between the 2 subscription groups ?Yes

Conclusion: **There was significant difference in performance of users who had trials subscription vs started subscription**

Is there a statistical difference in performance between users who have longer subscriptions ?

1. Is the total score performance significantly different between the 3 subscription duration group types? Yes.
2. Is the number of tests taken significantly different between the3 subscription duration group type ? Yes
3. Is the duration spent significantly different between the3 subscription duration group type ?Yes

Conclusion: **There was significant difference in performance of users who had subscription duration of 1 months, 3 months and 12 months.**



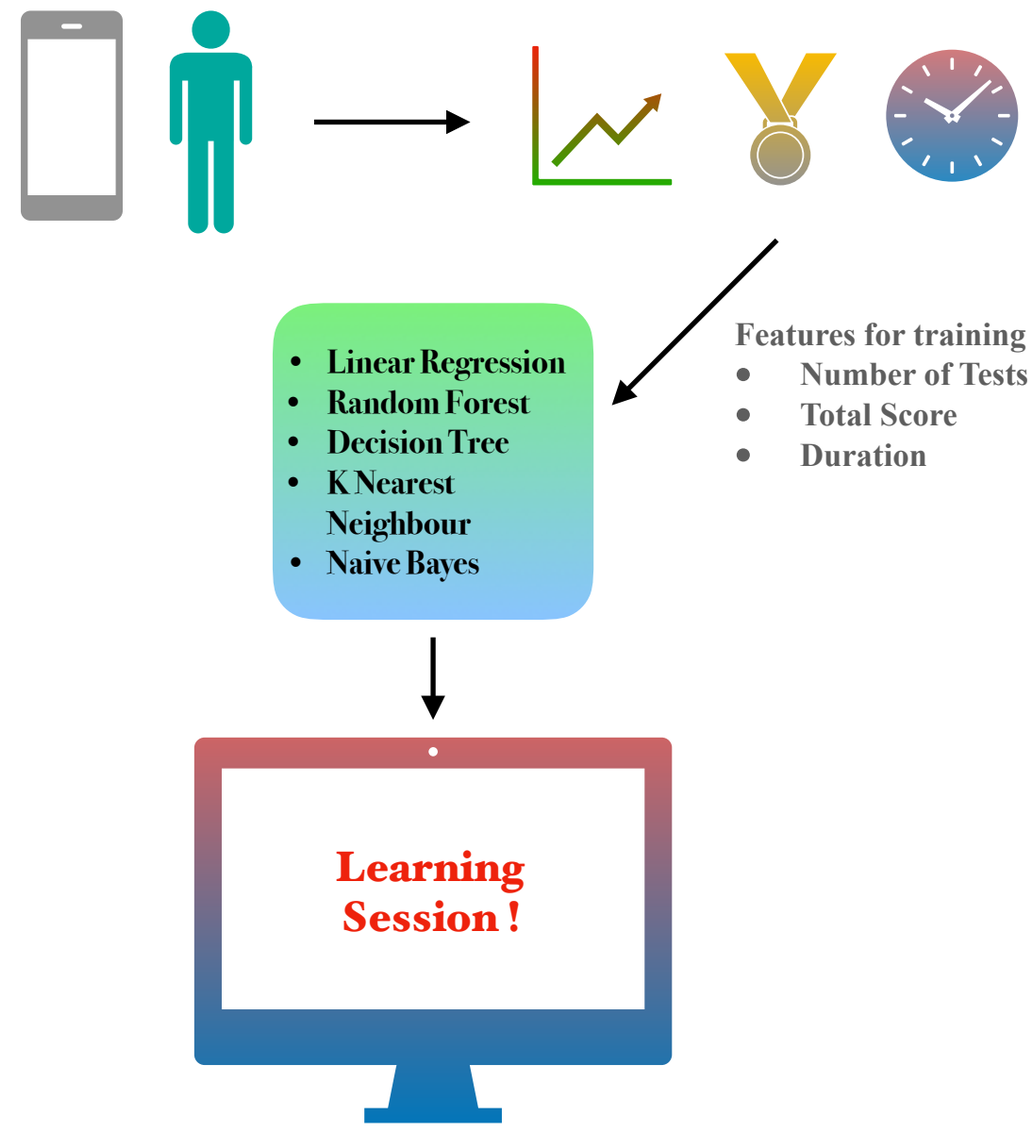
Modelling:

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 ***total scores obtained, number of tests done, and duration*** to predict and classify their learning type.

I have considered 2 approaches, for building this model.

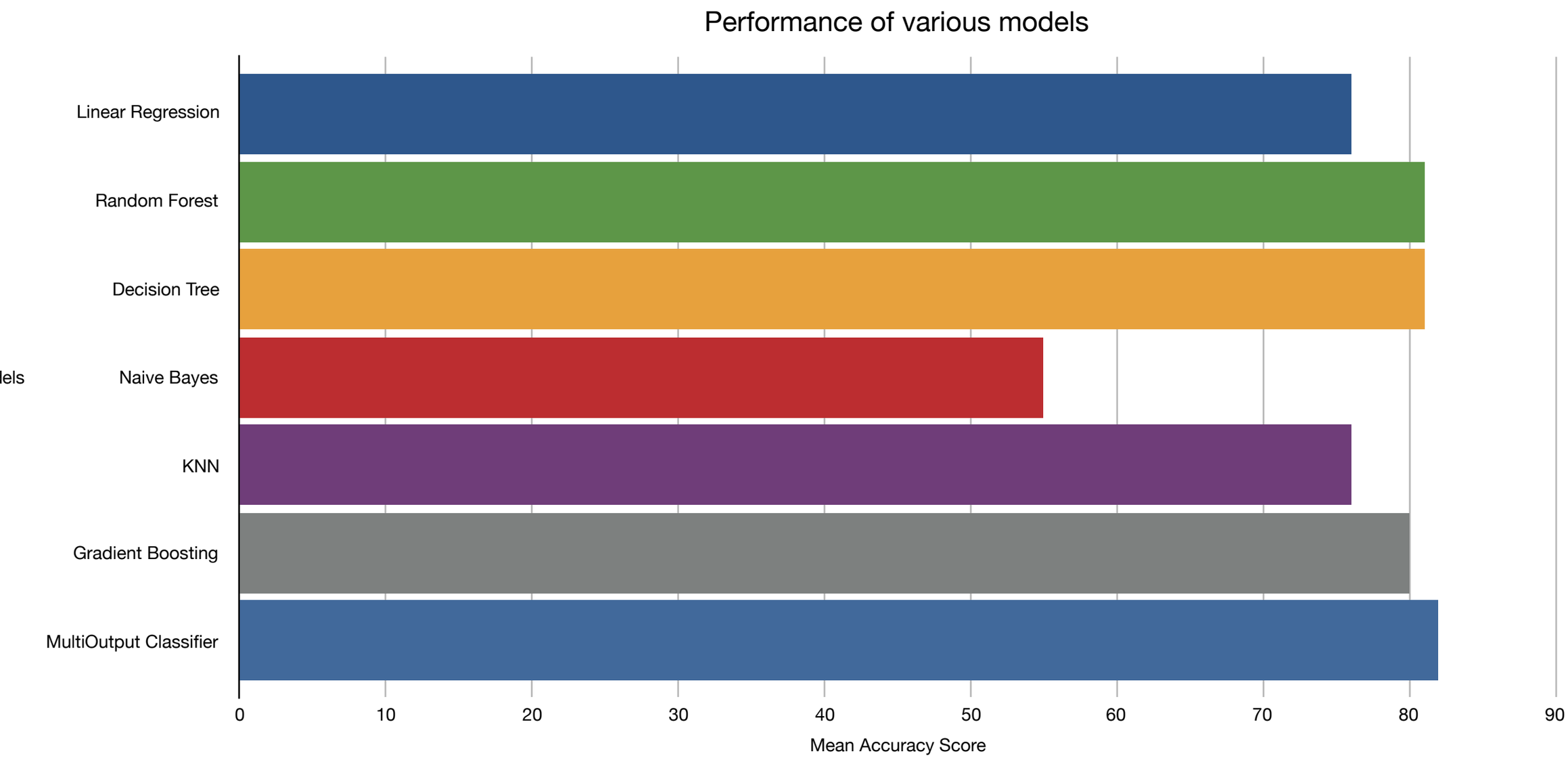
1. Use Classification techniques like Logistic Regression, Random Forest, Decision Tree, K nearest neighbours and Naive Bayes algorithms to train and test on the features and the target variable.
2. Ensemble approaches using advanced scikit learn algorithms like Gradient Tree Boosting, MultiOutput Classifier to improve the performance.

I have used the complete cleaned dataset to train and test the model of around 1.7million rows



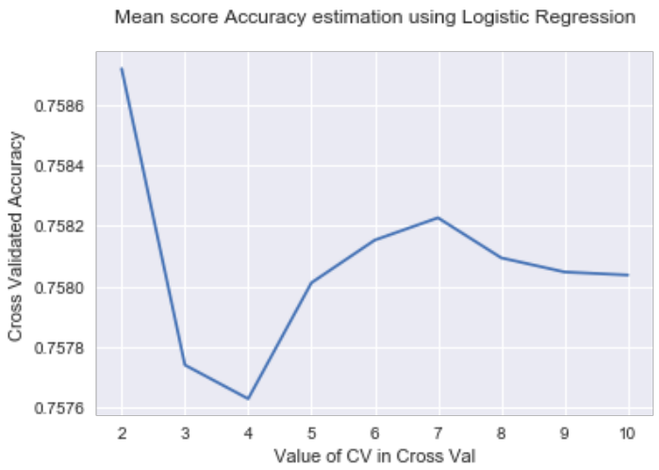
EMPATHIZE — IDEATE — PROTOTYPE — IMPLEMENT

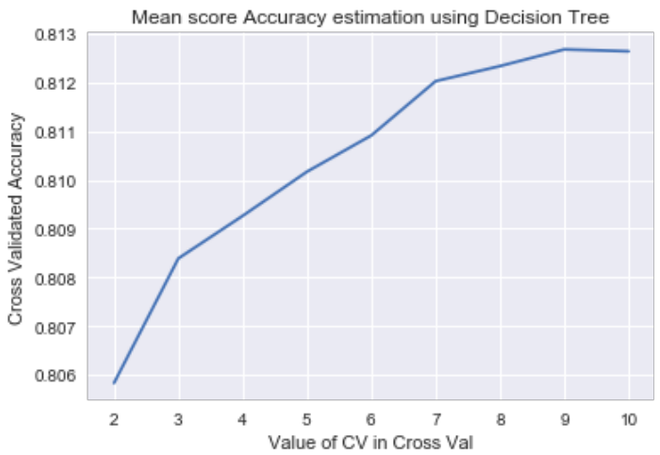
Performance Evaluation

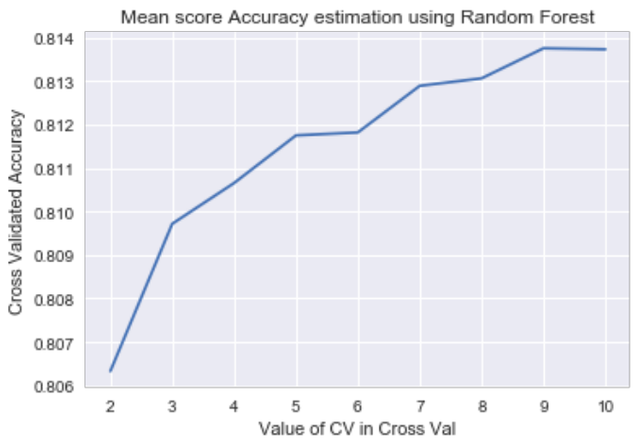


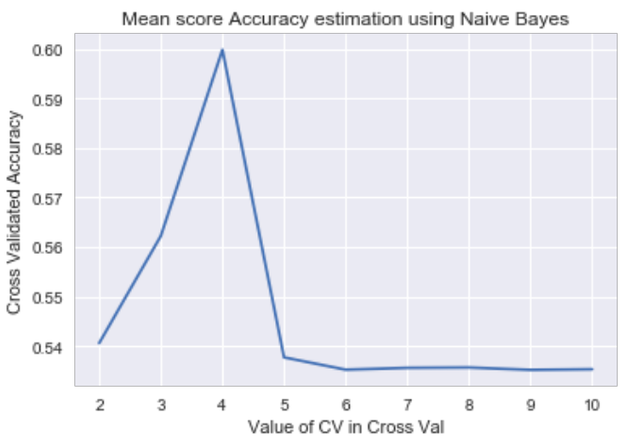
EMPATHIZE IDEATE PROTOTYPE IMPLEMENT

K Fold Cross Validation



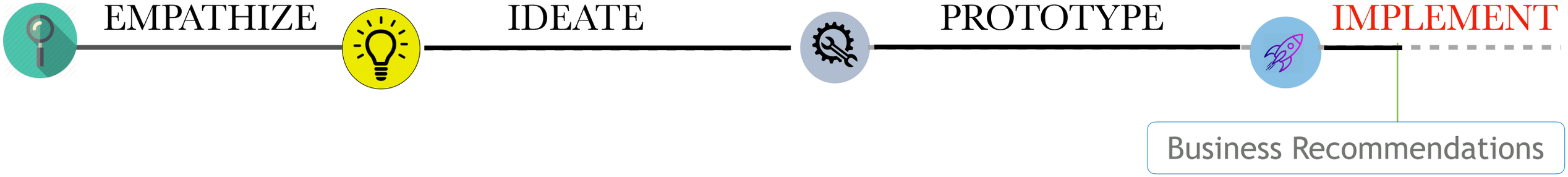






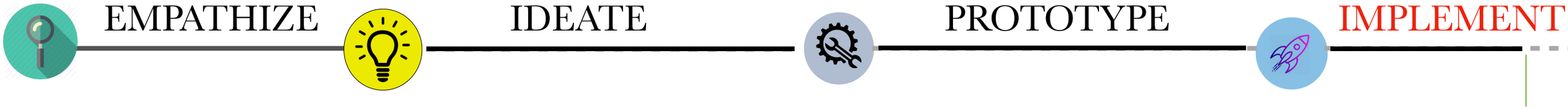
We are implementing K Cross Fold Validation to predict the mean accuracy of our models. Except Naive Bayes, the rest of the models have a very pretty good accuracy score to predict the user group. Ensembling does boost the performance of the model.

Overall the Random Forest, Decision Tree, and the ensemble approaches gave a 81% accuracy score to predict the different learning type of an user based on their total score, duration and number of tests taken. This is quite a good prediction score.



Business Recommendations and Use Cases:

1. Introduce new plans subscription plans for the countries that have low signup rates to motivate them and bring them into the platform.
2. Memrise web has a significant foothold in the top countries, and this should be targeted to the countries at the lower end of pyramid.
3. People spent a lot of time on the chat mission but yet the frequency of users in this phase is quite low. Needs to be understood what can be the reason for chat mission users being very low in number. Is the time or the mission UX ?
4. The average score is quite low and it shows the users do not have a very high scoring performance.



Future Scope:

1. Use unsupervised methods to understand customer behaviour and gain more insights.
2. Explore other 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. Use different encoding techniques to improve performance.
6. Other features by feature engineering can be explored to potential patterns and gain more insights about an user
7. Use Kernel Density estimation or Tukey test to remove identifiers for better cleaning of the dataset.

The Concluding Story

The Conclusion:

Over the course of the case study, I've gained some valuable insights to answer the following questions:

- How do the different data points relate to each other?

The different data points are related quite linearly as shown in the ideate phase, as an user who is doing more tests is prone to do better.

- Can you identify any groups of users based on their attributes? How do they differ?

In the ideate phase we have seen how different learning group users have significant difference in their performance,
I have tried to check the performance of the 3 different sets:

a. Set A: Groups based on different learning phases.
b. Set B: Groups based on different countries
c. Set C: Groups based on their subscription type

Also I was able to relate to performance comparison of different mediums (ie, Ios, Android, Web). Like Memrise web users were not existent in the least 5 countries by sign up frequency. Also all the learning sessions were not executed by these countries.

- Do some users, or groups, seem to have a better learning performance than others?

Groups that have done the learning phases like speed review, audio had better performances. Also web user groups had significant higher scores.

- Are you able to predict the accuracy of a learning session? How confident are you with the prediction?

Based on the algorithms, my model can predict unto 82% accuracy based on an users total score, number of tests done and duration in the app that which learning phase they are in.