Task completed:

| Date started | Date completed |
| --- | --- |
| 18.11.2020 | 23.12.2020 |

Analysis

Try and create 3 or more key success criteria for your program.

**Success Criteria:**

1. Analysing the VideoGame.csv data and testing multiple hypothesis, supported by visualization of the data.
2. Implementing a Multiple Linear Regression Model with Feature Selection and evaluating with chosen metrics. Metric values will be used to evaluate the success of our model.
3. Implementing a Random Forest Classifier which predicts a game being “Hit” or not with high recall and precision values.

Design

* *You may like to create a flow charts which will show broadly how your program will work. If so include your flow chart in this section.*
* *You must create pseudocode for a part of your program (minimum of 15 lines). If possible, try to create all of your program in pseudocode.*

1. Read and analyse Video Games Sales csv file. (Columns and data types)
2. Filter and clean the data according to determined limitations.
   1. Cleaning null values of columns that we will potentially use.
   2. Filter the data by looking the release dates and game quantities of each publisher.
3. Creating a new column for economical information.
   1. Research and determine what can represent a game’s economical power
   2. Get unique publisher names and research stock market revenue values of these companies.
   3. Prepare text file with names and links from [www.macrotrends.net](http://www.macrotrends.net)
   4. Implement a web scraper which takes the prepared text file as an input and gives a csv file as a result.
4. Preparing and cleaning GameCompRev data frame and merged with video games sales data frame.
5. Determine three different hypothesis and use visualization tools and Student T-Test to test the hypothesis.
6. Develop a Multiple Linear Regression model.
   1. Select from candidate features using their f scores.
   2. Creating the model with transformed train data.
   3. Prediction and evaluating the prediction with chosen metrics.
7. Develop a Random Forest Classifier model in order to predict the games with over 1 million global sales value.
   1. Creating a isHit function in order to label games with over 1 million global sales value with 1
   2. To use the categorical columns in our model create their dummies.
   3. Creating the model with transformed train data.
   4. Prediction and evaluating the prediction with chosen metrics.

**Linear Regression Model Pseudocode:**

\*Select the necessary columns with candidate features and determined label.

\*Separate label and feature

\*Scale the numeric columns with StandardScaler()

\*Separate Train and Test data

\*Feature selection algorithm with SelectKBest and f\_regression

\*Visualization of feature F Scores

\*Creating and feeding the model with prepared data

\*Prediction with the X\_test data

\*Evaluation of the results:

fMSE = metrics.mean\_squared\_error(y\_test,y\_pred\_f)

fMAE=metrics.mean\_absolute\_error(y\_test, y\_pred\_f)

fRMSE = np.sqrt(metrics.mean\_squared\_error(y\_test,y\_pred\_f))

**Pseudo code for Predicting Hit Games with RFC:**

Create new data frame with existing data frame

Create new column named ‘Hit’ with ‘Global\_Sales’ column values.

Drop ‘Global\_Sales’ column

def isHit(sale):

if sale equal and bigger than 1

return 1

else

return 0

Use isHit function for every row in ‘Hit’ column

Create new data frame with existing one and apply get\_dummies function

Choose labels and features

Apply train\_test\_split function in order to create test and train data

Apply RandomForestClassifier function in order to get predicted values.

**My tests:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | What am I testing? | What data will I use? | Normal/Boundary/Erroneous? | Expected Result |
| 1 | We tested to select the most suitable feature with f\_regression for my model. | Normal Data | Normal test | Finding the most suitable at least 4 features for my model. |
| 2 | We made prediction about which are hit games(global sale is more than 1 milion) and we tested their hit prediction. | Normal Data | Normal Test | High percentage of hit precision. |
| 3 | We made Student T-Test to accept or reject our initial hypothesis. | Normal Data | Normal Test | We expected that P value of T-Test is less than 0.05 for. |

Development

* *Copy and paste your code into this section(you can also write github, drive or dropbox link of your project instead of putting project codes here)*
* *Remember to try and add comments to your code to make it more readable!*

**My program code:**

<https://github.com/korhankoz/AdvancedPythonProject>

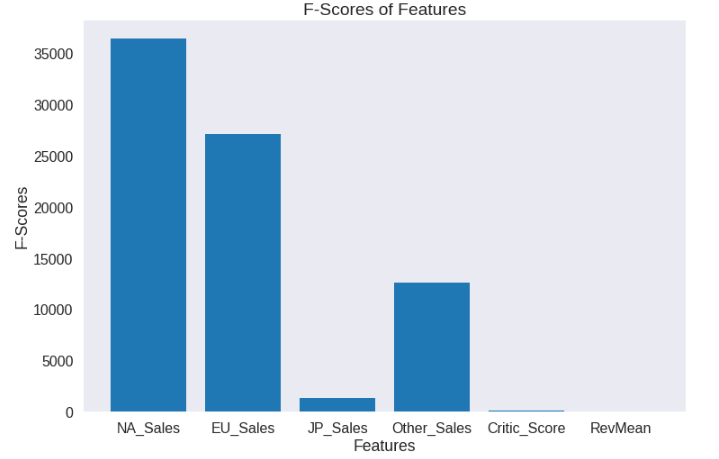
Testing

Since our project is a data mining project, no system testing was done.But instead, we did modal testing, which we tested and controlled its stages.These tests are as below.

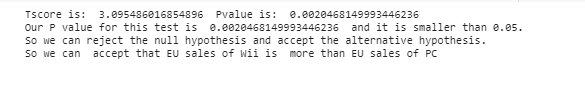
**My tests:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | What am I testing? | Expected result | Pass/Fail | Do I need to change my program? If so, how? |
| 1 | We tested to select the most suitable feature with f\_regression for my model | Finding the most suitable at least 4 features for my model | 3 valuable features were found, but the 4th value was not evaluated because of it was not suitable. | We will transform our initial train data according to the features we selected.So we will use these features in the regression part. |
| 2 | We made prediction about which are hit games(global sale is more than 1 milion) and we tested their hit prediction. | High percentage of hit precision | Recall value for 0’s prediction is 0.92.This shows success.  Recall value of 1’s is 0.48. this recall value could be higher and improved. | We can add more numeric features. |
| 3 | We made Student T-Test to accept or reject our initial hypothesis. | We expected that P value of T-Test is less than 0.05 for accepting our hypothesis. | This test passed.Our P value is less than 0.05. | No we do not..Because our test is passed and it means we can accept the hypothesis. |

**My test screenshots:**



**For Test -1**



**For Test -2**

**For Test -3**

Evaluation

* Evaluate how successful your program was. You should like your evaluation to your testing results.
* You should reflect on any new skills you have developed

This section should be approximately 200-500 words.

**How successful was my program?**

Success of a Data Science project can be evaluated by many aspects,

We used proper visualization tools for desired tables and commented about what the data we worked on can tell us. We used two different Machine Learning Models which were Linear Regression and Random Forest Classifier.

Implementation didn’t have any flaws but the results and the features that we used when feeding the models can be evaluated furthermore. We added a new column to our data which presented information of economical power of the publisher firm. We created that column by using mean revenue values of the companies from the stock market. Eventually we saw that we can not use this column as a feature by looking at its F regression score. This has a simple explanation and solution. We could not find the exact amounts that companies spent on the game departments so that economical indicator represented economical power of the company not the game. Solution for this is simple but was not achievable by us, because we didn’t have any access to detailed economical information of the companies.

Our idea and research can be evaluated as successful, but we had our limits when it comes to finding data about the subject. Usage of web scraping was successful too, we wanted to demonstrate that part because it is a skill we learned during our course but out of the Data Science subject.

Also, in Random Forest Classifier part our model was not so successful. The reason of this, when trying to predict global sales values for games, we did not use numerical columns that could be related to global sales values such as NA\_Sales or EU\_Sales columns. Probably that’s why our precision and recall scores for 1 is smaller than 0. “

**What new skills have I developed?**

Using web scraping was new for us, also we use all the new skills we learned during the course under the Data Science Subject. Visualization tools and their customization, implementing feature selection with F scores and etc. We sharpened the skills we already learned and gained new ones