Text-mining Assignment Submission Cover Sheet

This Assessment Cover Sheet **must** be included on all Assessment submissions.

|  |  |
| --- | --- |
| Assignment Title | CA2 Data Mining – Predicting Medication Rating |
| Module | DATA9900: 2022-23 |
| Student Name  (same as Student Card) | Maurício de Oliveira Reis |
| Student Number | D21125621 |
| Programme | TU256/1 |
| Part-Time/Full-Time | Part-Time |
| Year of Study  (First Year, Second Year, etc) | 2022/2023 |

Late Submissions: Assessment submitted after the deadline will have a late penalty applied.

**Academic Integrity for assessment in TU Dublin Programmes**

Each student is responsible for knowing and abiding by TU Dublin Academic Regulations and Policies. Any student in breach of these regulation/policies will be subject to action in accordance with the University’s procedures for breaches of assessment regulations. Please refer to the General Assessment Regulations at

<https://tudublin.libguides.com/c.php?g=674049&p=4794713>

<https://www.tudublinsu.ie/advice/exams/breachesofregulations/>

All students are expected to complete their courses/programmes in compliance with University regulations. No student shall engage in any activity that involves attempting to receive a grade by means other than honest effort, for example:

1. No student shall complete, in part or in total, any examination or assessment for another person.
2. No student shall knowingly allow any examination or assessment to be completed, in part or in total, for themselves by another person.
3. No student shall plagiarise or copy the work of another and submit it as their own work.
4. No student shall falsify any data. Falsification is the invention of data, its alteration, its copying from any other source, or otherwise obtaining it by unfair means, or inventing quotations and/or references.
5. No student shall use aids or devices excluded by the lecturer in undertaking course work or assessments/ examinations.
6. No student shall knowingly procure, provide, or accept any materials that contain questions or answers to any examination or assessment to be given at a subsequent time.
7. No student shall provide their assignments, in part or in total, to any other student in current or future classes of this module/ programme unless authorised to do so by the lecturer.
8. No student shall submit substantially the same material in more than one module/programme without prior authorization.
9. No student shall alter graded assignments or examinations and then resubmit them for regrading, unless specifically authorised to do so by the lecturer.
10. All programming code and documentation, unless correctly referenced, submitted for assessment or existing in the student’s computer accounts must be the students’ original work or material specifically authorized by the lecturer.
11. Collaborating with other students to develop, complete or correct course work is limited to activities explicitly authorized by the lecturer.
12. For all group assignments, each member of the group is responsible for the academic integrity of the entire submission. Consequently, all group members must satisfy themselves that all elements of their submission adhere to the academic integrity statement points above.

By submitting coursework, either physically or electronically, you are confirming that it is your own work (or, in the case of a group submission, that it is the result of joint work undertaken by members of the group that you represent) and that you have read and understand the University’s Regulations and Policies covering Academic Integrity (see General Assessment Regulations)*.*

Coursework may be submitted to an electronic detection system in order to help ascertain if any plagiarised material is present. If you have queries about what constitutes plagiarism, please speak to your lecturer.

|  |  |
| --- | --- |
| Student Signature |  |
| Date | 28/12/2022 |

IMPORTANT:

* Sections 1-8 should be no longer than 15 pages (minimum 10 pages), including diagrams, images, screen captures, tables, etc. Careful selection of these is needed.
  + Code does not count to this total. Code should be added to the relevant section.
  + Detailed discussion is expected. Marks are awarded based on depth of information given.
  + You can add as many references as needed. They are not included in the limit of pages.

1. **Definition of Problem**

Producing new drugs is expensive, a study in 2020 estimated that the median cost of getting a new drug into the market was $985 million (Cost of drug development , 2022).

Because of that, while in development and test, it may be relevant to consider not only the effectiveness of the drug but also have an idea of how well rated this drug would be by the people that will use it. Here we will look at many attributes of the Drug Review Dataset (Drug Review Dataset (Druglib.com) Data Set, 2018), and try to understand better what makes a drug be well rated by its users. Insights from this work can lead not only to cheaper drugs for patients but even new drugs on the market that would be considered economically unviable before depending on how this work might influence industry guidelines. It might even prevent some economically unviable drugs from being produced.

1. **Data Exploration & Descriptive Analytics**

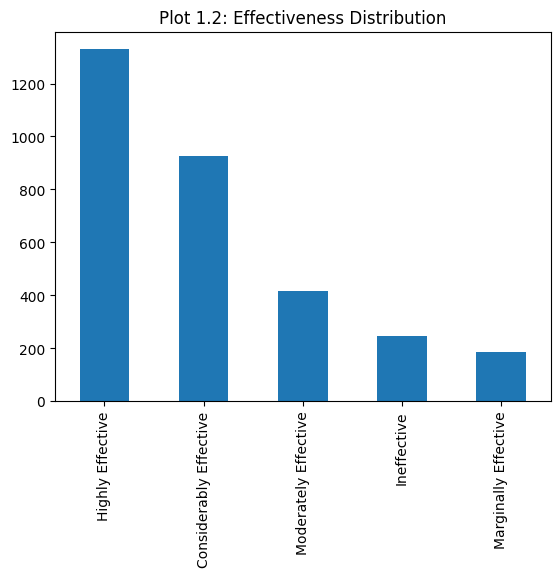
Our target variable here will be rating, as we can see on Plot 1, this is a left skewed distribution with values between 1 and 10. Although it looks more like a left skewed distribution, it has a somewhat large concentration of rating 1 showing that it’s more common to either give a rating or 7 or above or a 1 for a medication.

Chart, histogram

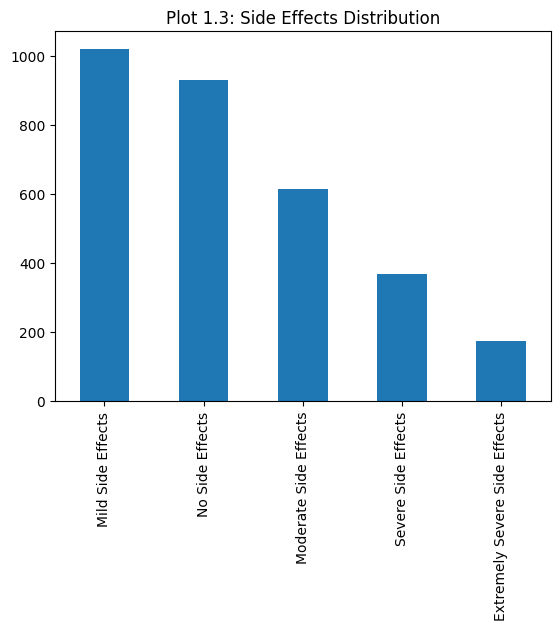
Description automatically generated

Apart from rating, for our analysis we will consider the following variables:

- **effectiveness**: categorical variable with 5 levels that goes from Highly Effective to Ineffective. Most drugs are highly or considerably effective as shown on Plot 1.2



- **sideEffects**: categorical variable with 5 levels that goes from Extremely Severe to No Side Effects. Most drugs have mild or no side effects as shown on Plot 1.3



- **benefitsReview**: a free text field where patients put a review of the medication benefits. We will apply a sentiment analysis to understand if this is positive or negative review and create a new field called benefitsReviewSentiment for our analysis.

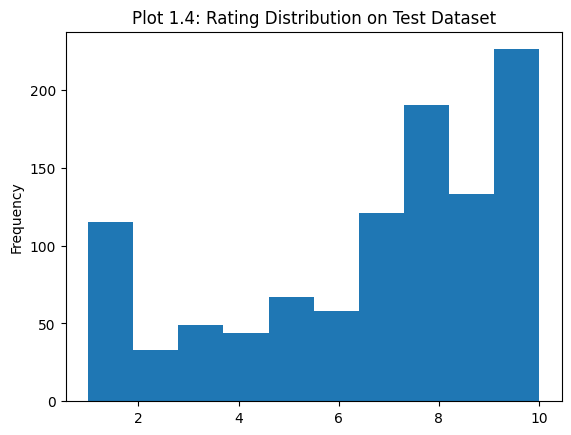
- **sideEffectsReview**: a free text field where patients put a review of the medication side effectgs. We will apply a sentiment analysis to understand if this is positive or negative review and create a new field called sideEffectsReviewSentiment for our analysis.

- **commentsReview**: a free text field where patients put a review of the medication in general. We will apply a sentiment analysis to understand if this is positive or negative review and create a new field called commentsReviewSentiment for our analysis.

As we will look at overall rating of medications and not each individual drug, we will not consider on this analysis the urlDrugName, condition will also not be considered as it is a categorical variable with nearly as many different levels as urlDrugName and, unless somewhat grouped, wouldn’t help on this specific analysis. Also, we will not consider the first unnamed column since it doesn’t have a description of what it is on the datasource website.

**Rating distribution on test dataset**

We also looked at the rating distribution on the test dataset on plot 1.4 to see if it has roughly similar distribution that the train dataset to access if the data was spitted properly



As we can see, plot 1.4 has a roughly similar shape to plot 1.

1. **Data Preparation**

We’ve transformed effectiveness and sideEffects into categorical variables.

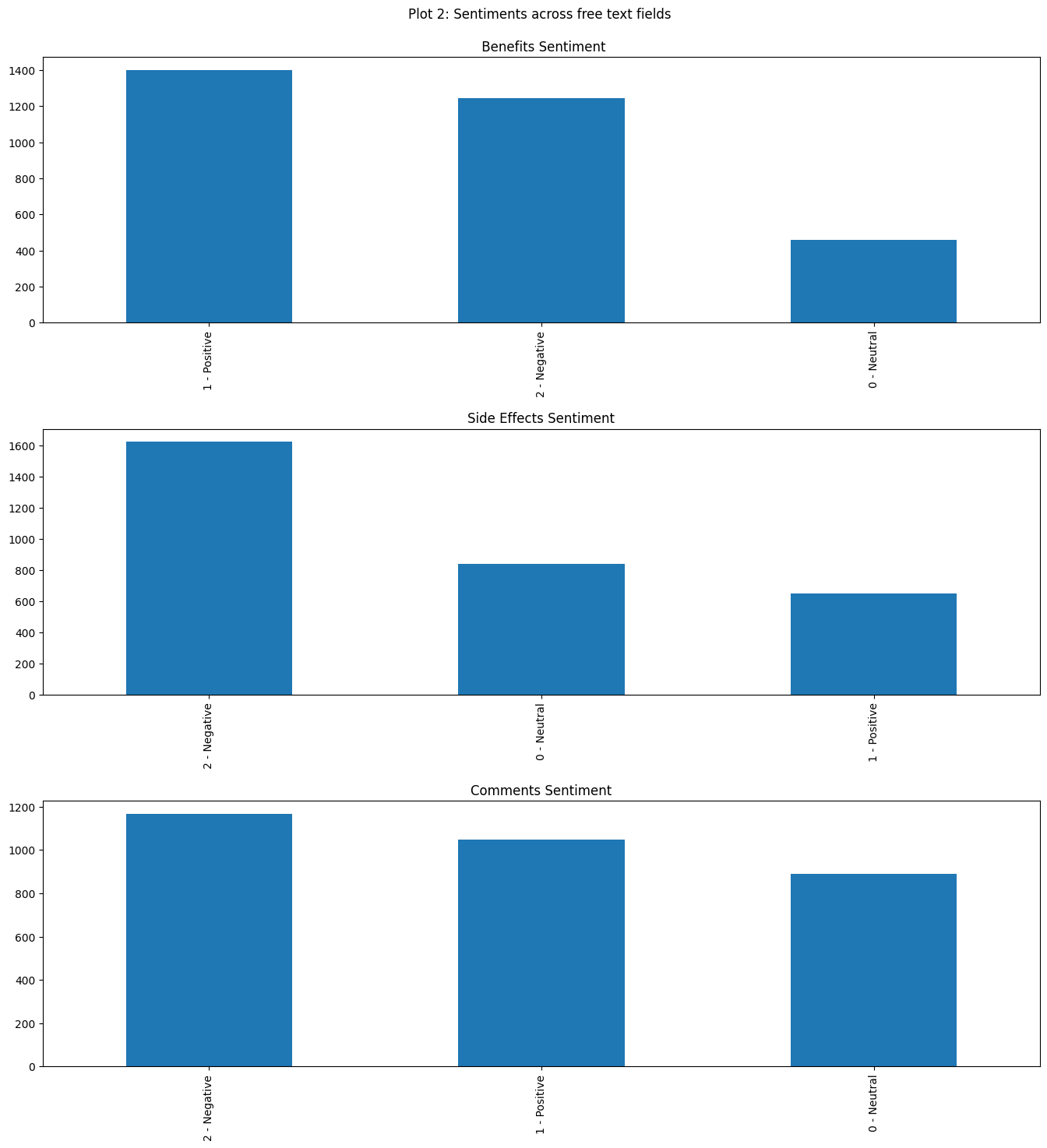
We’ve also checked for NAs, which revealed a very small quantity of NA values for sideEffectsReview 0.06% of its rows and commentsReview 0.26% of its rows. As the number is small, we’ve decided to just consider the sentiment of those reviews as neutral.

**Sentiment analysis**

The Sentiment analysis was performed in the 3 reviews fields of the dataset which were free text fields: benefits, side effects and comments review.

We transformed those fields in String and made all its text lower case. After that, we’ve used the lib Vader Sentiment (VADER-Sentiment-Analysis, 2022) to decide whether the text contained on each review as positive, negative or neutral.

On plot 2 we can see that most sentiment on benefits field were positive where in the side effects field were negative. We can also see that the comments section has more negative comments.



1. **Details of Algorithms & Configurations**

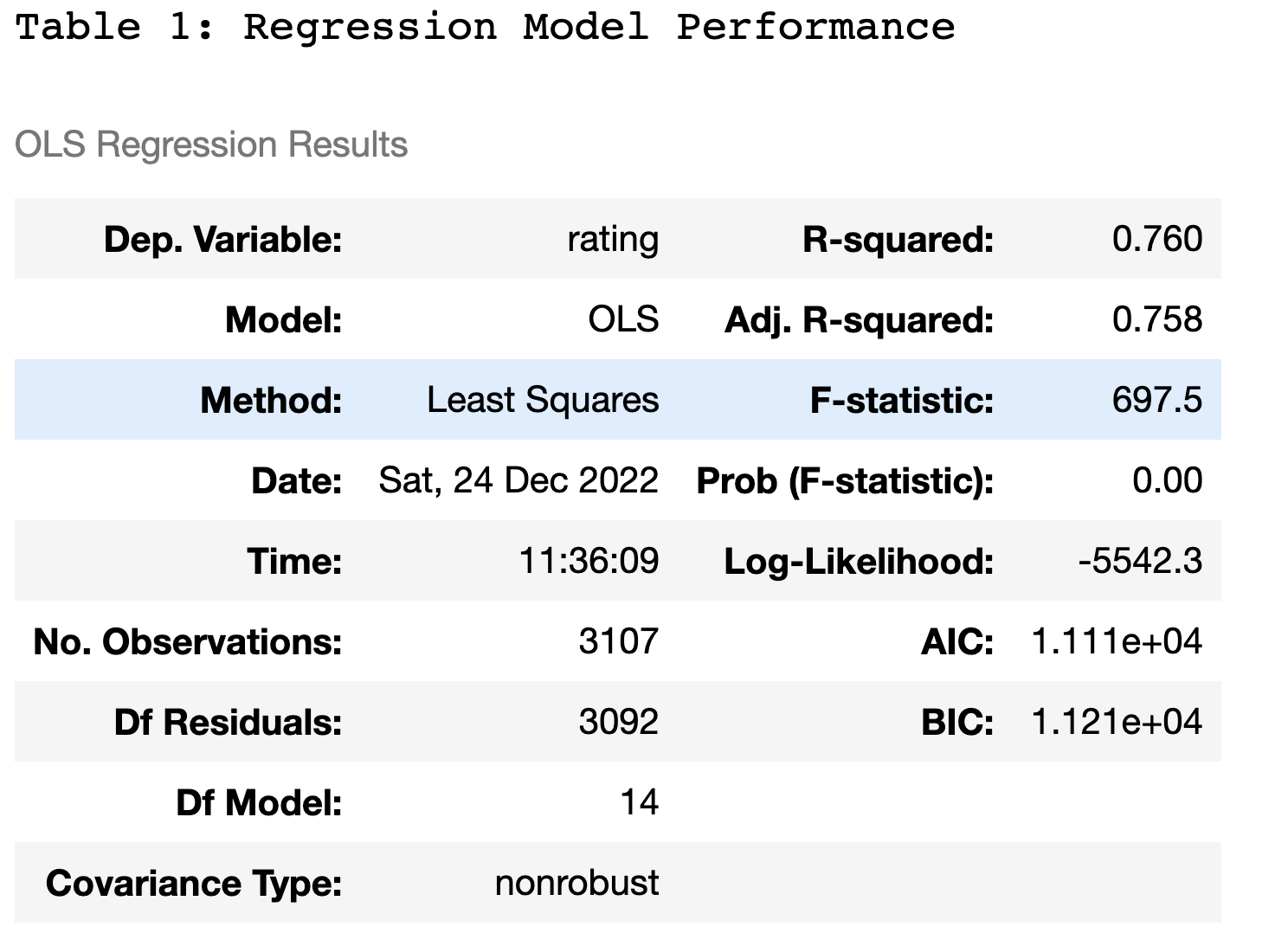
The algorithm utilized was a Linear Regression algorithm, for this we’ve utilized the following formula:

rating ~ C(effectiveness) + C(sideEffects) + C(benefitsSentCat) + C(sideEffectsSentCat) + C(commentsSentCat)

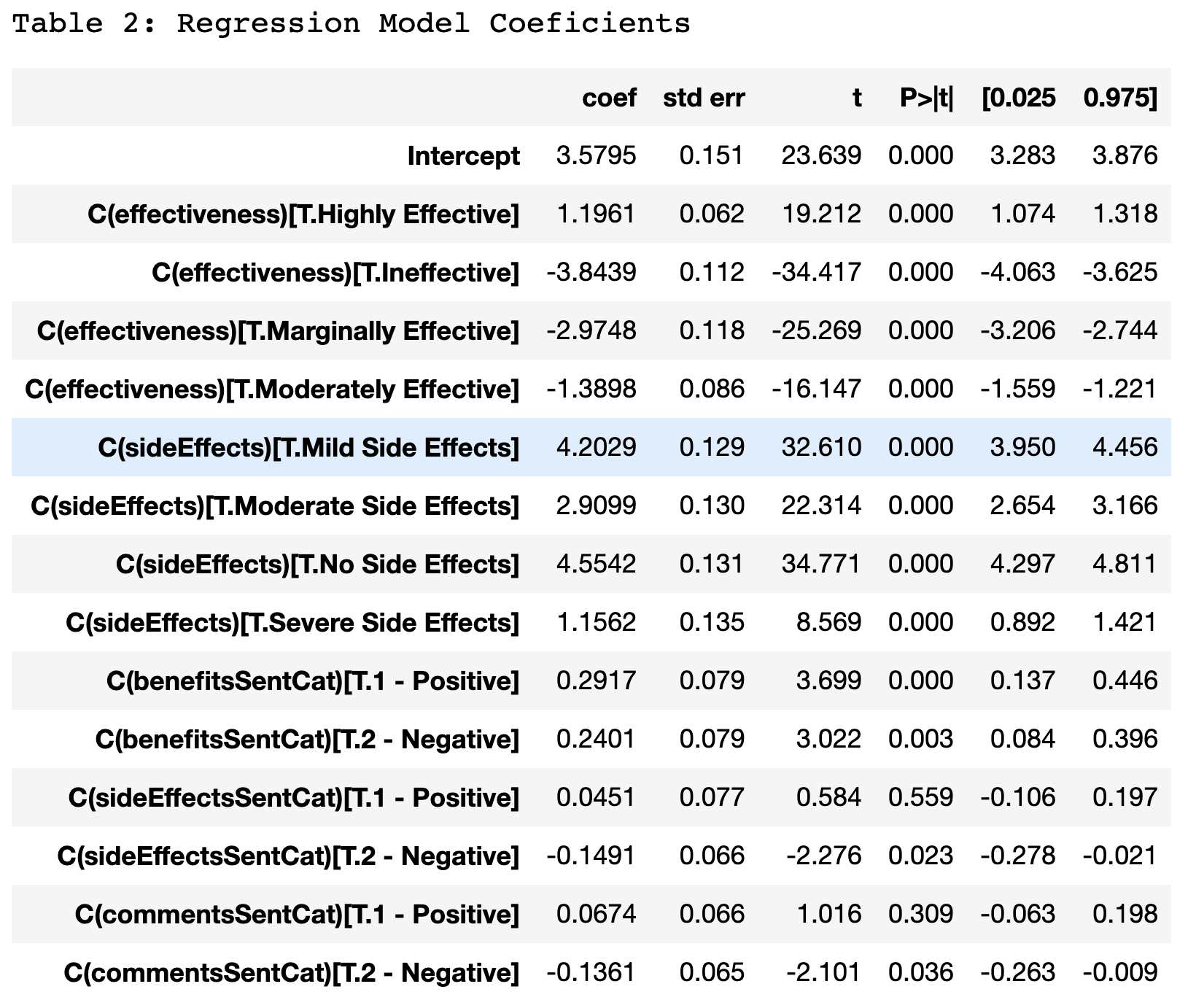
This reads as “rating explained by effectiveness, sideEffects, benefitsSentCat, sideEffectsSentCat and commentsSentCat”

The library utilized to build this model was the statsmodels (statsmodels, 2022) , we’ve utilized the subpackage statsmodels.formula.api which allows to use R like syntax to build models and facilitate the handling of categorical variables.

1. **Model Performance Metrics & Evaluation of Results**



On Table 1 we can see several attributes about our model, the adjusted R Squared tells us that this model is capable of explain 75.8% of the variance from rating.

****

On Table 2 we can see the coefficients for this regression model. To avoid the dummy variable trap, “Considerably Effective” effectiveness, “Extremely Severe Side Effects” sideEffects, and the neutral benefits sideEffects and comments reviews values were aggregated all aggregated to the Intercept which is 3.5795.

We can only have one coefficient counting for each one of the predictor variables so, if we count the + 0.0674 of commentsSentCat ->Positive, the other coefficient of -0.1361 from the negative value wouldn’t count to calculate the expected score. So, we can expect that a medication that is ineffective, has no side effects and is reviewed positively on the benefits, side effects and comments sections” would have the following score:

3.5795 – 3.8439 + 4.5542 + 0.2917 + 0.0451 + 0.0674 = **4.694**

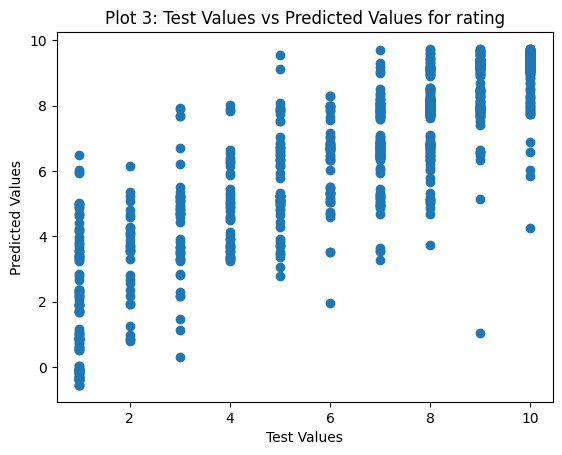
On Table 2 we can also see what contributes the most to a medication score. We can see that side effects and effectiveness are by far the attributes of a medication that contribute the most for its score. Although, there is a huge difference between Highly effective and considerably effective (1.1961) whereas there is a much smaller difference between No Side effects and Moderate Side effects (4.5542-4.2029=0.3513) suggesting for example that people don’t mind that much some side effects from a medication but mind more about a small change in the level of effectiveness of this medication.

**Using test data to access model performance**

To access the model performance in practice, we’ve utilized the test data

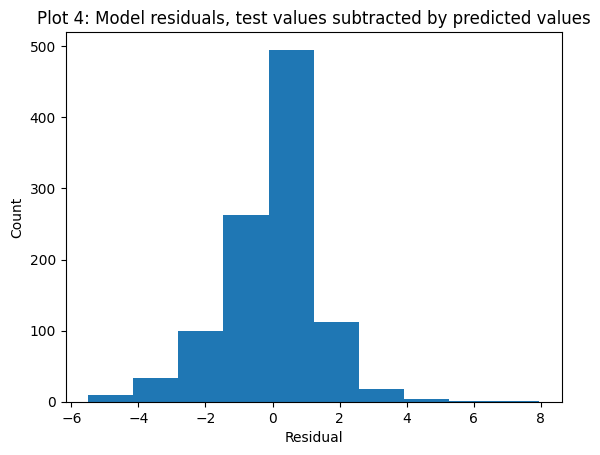
that comes with the Drug Review Dataset mentioned before that we’ve utilized to build this model. The test dataset has the same columns as the train one, no NA values and we performed the same data transformations that we performed on the train dataset to define categorical variables and sentiment analysis for the free text fields.

Once the transformations were done, we used the data from the test dataset to predict rating values and compared those values with the original values present on the test dataset.

****

As we can see on the scatterplot on plot 3, there a correlation between the rating values predicted by the model and the rating values present on the test dataset.

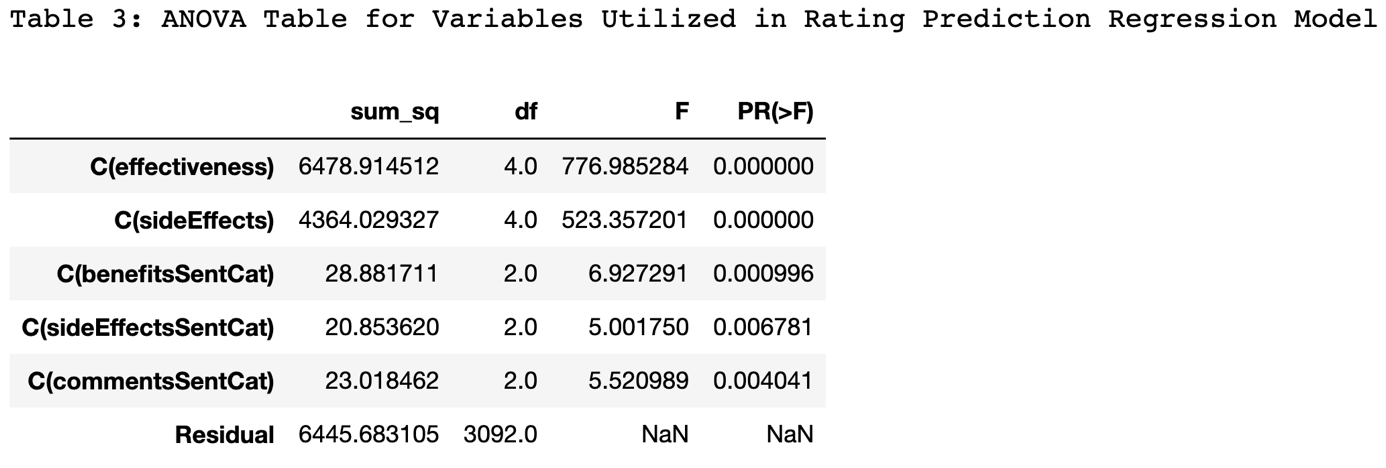
Also, to further evaluate this model, you can see on plot 4 the residuals plot.



The residuals are the result of the subtraction between the test and the predicted rating results. The more normal this plot looks like, the more adequate the model is. Here we have a nearly normal looking distribution suggesting that this model is not that bad in predicting rating values.

1. **Identification of the most important variables**

To identify which variables were important to build our model, we’ve utilized an ANOVA table. As we can see on table 3, all variables were considered significant (*p* < 0.05). Therefore we’ve concluded that all those variables influence on the rating of a medication.



1. **Comparison with other Research & Reflections**

**Predicting rating based on reviews** (Mohamed, 2018)

This Kaggle notebook utilized a neural network from TensorFlow (tensorflow, 2022) to predict rating based on user reviews. The work focus on measuring the overall accuracy and loss of the model, it’s a different approach from what we saw on this assignment with a different goal as here the focus is more to identify what can impact on user rating, predicting it exactly is not a priority. The most interesting part of this work on Kaggle is that they grouped ratings since some groups had very low representativity, that could have been done here on this assignment as well.

**Rating prediction project** (Cheng, 2019)

This Kaggle notebook utilized naïve bayes to predict rating. Although it didn’t focus on the particular effect of each variable again, its testing accuracy of 0.876 was quite high compared to other works and wit this assignment work as well. The most interesting part of this work for me was to see the power and the simplicity of Naïve Bayes when all you want is a prediction task, also was interesting to see how the overfitting issues were handled there by grouping the ratings.

**Predicting Drug Ratings** (Roznovjak, 2018)

This model also utilized a neural network form TensorFlow (tensorflow, 2022) to predict rating. Apart from being very detailed on the work, not only showing code like most of other Kaggle notebooks based on this dataset, this work looks deep into how the reviews text are formed and which special characters/html tags could get in the way of the sentiment analysis. Also, it mentions in the beginning that it didn’t do an exploratory analysis because there were many already available on the Internet that he could base his work on, something that I would definitely do in the future when working with a popular dataset like this one (probably not a very common everyday situation though).

1. **Action plan to improve equality and other social issues based on your data**

This work looked at what matters most to patients in a drug for them to give a good rating for this drug. Although rating probably shouldn’t be used as the only guideline to produce a new drug, this work can help to focus on what matters most to make a drug viable marketwise by giving insights like for example “the difference between no side effects and mild ones don’t affect much the user rating on a drug”. Insights like this can help pharmaceutical companies to focus on what matters most and, with that, reduce the production cost of certain drugs which would ultimately impact patients who have to pay for those drugs.

1. **References**

# Bibliography

Cheng, H. (2019, 12 28). *rating prediction project* . Retrieved from kaggle: https://www.kaggle.com/code/harrycheng5/rating-prediction-project

Mohamed, M. (2018, 12 28). *Predecting rating based on reviews* . Retrieved from kaggle: https://www.kaggle.com/code/moghazy/predecting-rating-based-on-reviews

Roznovjak, M. (2018, 12 28). *Predicting Drug Ratings* . Retrieved from kaggle: https://www.kaggle.com/code/avatust/predicting-drug-ratings

statsmodels. (2022, Nov 02). *statsmodels*. Retrieved from statsmodels: https://www.statsmodels.org/stable/index.html

tensorflow. (2022, 01 21). *tensorflow*. Retrieved from tensorflow: https://www.tensorflow.org/

UCI. (2018, 10 02). *Drug Review Dataset (Druglib.com) Data Set*. Retrieved from uci: https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Druglib.com%29#

VADER-Sentiment-Analysis. (2022, April 1). *vaderSentiment*. Retrieved from github: https://github.com/cjhutto/vaderSentiment

Wikipedia. (2022, November 29). *Cost of drug development* . Retrieved from wikipedia: https://en.wikipedia.org/wiki/Cost\_of\_drug\_development