Data Intake Report - Data Science Healthcare - Persistency of a Drug - Classification

Group Name: <Igor Azevedo de Queiroz>

Email: <[igor\_queiroz17@yahoo.com.br](mailto:igor_queiroz17@yahoo.com.br)>

Country: <Ireland>

College: <Dublin Business School>

Specialization: <Data Science - Classification>

Report date: <20/09/2021>

Internship Batch:<LISUM02>

Version:<1.0>

Data storage location: <https://github.com/IgorQueiroz23031988/Taxi-Cab-Data-Analysis-and-Insights>

Problem Description: <Data Science Healthcare - Persistency of a Drug - Classification>

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | < 3424rows > |
| **Total number of files** | <1> |
| **Total number of features** | < 69 columns > |
| **Base format of the file** | <.ipynb, .csv, .txt, .png> |
| **Size of the data** | <891 in KB> |

**Healthcare - Persistency of a Drug - Classification**

**1. Business Description/ Problem.**

One of the challenges for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

With an objective to gather insights on the factors that are impacting the persistency, it is necessary to build a classification for the given dataset, using the variable ‘Persistency\_Flag’ as target variable and other attributes as prediction variables.

**2. Business Understanding.**

ABC it is a private pharma company. Due to the problem to the persistency of drug as per the physician prescription, a data science project is applied to predict the classification of ‘Persistency\_Flag’ variable. In other words, based on the previously patients characteristics it is possible predict if futures patients will use the drugs during the role treatment or if they won’t.

The object of this project is providing answer of the main questions made by the company’s CEO, which are:

* What is the ‘Persistency\_Flag’ classification for future patients?

The answer for those questions is presented in two different methods:

* A webapp with all necessary prediction attributes in order to predict the classification of the ‘Persistency\_Flag’ for future patients.
* A dashboard with several hypotheses and insights to help the company CEO with future decisions.O decisions.

The tools used for this project are: Python 3.8, Pycharm, Jupyter Notebook, Streamlit and Heroku.

**2. Data Understanding.**

There is 1 dataset provided:

<https://www.kaggle.com/harbhajansingh21/persistent-vs-nonpersistent>

healthcare\_dataset.csv – this file includes characteristics of several patients.

Variables Description:

Here I'm describing the columns in detail:

Patient Details:

* **Patient ID:** Unique ID of each patient;
* **Persistency\_Flag:** Flag indicating if a patient was persistent or not;
* **Age:** Age of the patient during their therapy;
* **Race:** Race of the patient from the patient table;
* **Region:** Region of the patient from the patient table;
* **Ethnicity:** Ethnicity of the patient from the patient table;
* **Gender:** Gender of the patient from the patient table;
* **IDN Indicator:** Flag indicating patients mapped to IDN;

Provider Attributes:

* **NTM - Physician Specialty:** Specialty of the HCP that prescribed the NTM Rx;

Clinical Factors:

* **NTM - T-Score:** T Score of the patient at the time of the NTM Rx (within 2 years prior from rxdate);
* **Change in T Score:** Change in Tscore before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown);
* **NTM - Risk Segment:** Risk Segment of the patient at the time of the NTM Rx (within 2 years days prior from rxdate);
* **Change in Risk Segment:** Change in Risk Segment before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown);
* **NTM - Multiple Risk Factors:** Flag indicating if patient falls under multiple risk category (having more than 1 risk) at the time of the NTM Rx (within 365 days prior from rxdate);
* **NTM - Dexa Scan Frequency:** Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate);
* **NTM - Dexa Scan Recency:** Flag indicating the presence of Dexa Scan before the NTM Rx (within 2 years prior from rxdate or between their first Rx and Switched Rx; whichever is smaller and applicable);
* **Dexa During Therapy:** Flag indicating if the patient had a Dexa Scan during their first continuous therapy;
* **NTM - Fragility Fracture Recency:** Flag indicating if the patient had a recent fragility fracture (within 365 days prior from rxdate);
* **Fragility Fracture During Therapy:** Flag indicating if the patient had fragility fracture during their first continuous therapy;
* **NTM - Glucocorticoid Recency:** Flag indicating usage of Glucocorticoids (>=7.5mg strength) in the one year look-back from the first NTM Rx;
* **Glucocorticoid During Therapy:** Flag indicating if the patient had a Glucocorticoid usage during the first continuous therapy;

Disease/Treatment Factors:

* **NTM - Injectable Experience:** Flag indicating any injectable drug usage in the recent 12 months before the NTM OP Rx;
* **NTM - Risk Factors:** Risk Factors that the patient is falling into. For chronic Risk Factors complete lookback to be applied and for non-chronic Risk Factors, one year lookback from the date of first OP Rx;
* **NTM - Comorbidity:** Comorbidities are divided into two main categories - Acute and chronic, based on the ICD codes. For chronic disease we are taking complete look back from the first Rx date of NTM therapy and for acute diseases, time period before the NTM OP Rx with one year lookback has been applied;
* **NTM - Concomitancy:** Concomitant drugs recorded prior to starting with a therapy(within 365 days prior from first rxdate)  
  **Adherence:** Adherence for the therapies.

**3. Business Assumptions.**

For further analysis, it is necessary to know the period of time of each season on US, those information will be included in a variable called Season. According to the website: <https://www.timeanddate.com/calendar/aboutseasons.html>, each season is specified on the followed days:

* spring runs from March 1 to May 31;
* summer runs from June 1 to August 31;
* fall (autumn) runs from September 1 to November 30; and
* winter runs from December 1 to February 28 (February 29 in a [leap year](https://www.timeanddate.com/date/leapyear.html)).

Due the dataset City.csv does not inform the State of two cities: SILICON VALLEY and ORANGE COUNTY, it is necessary specify the States of them.

According to the website: <https://www.britannica.com/place/Silicon-Valley-region-California>, The SILICON VALLEY State is California, CA.

Also according to the website: https://en.wikipedia.org/wiki/Orange\_County,\_California, The ORANGE COUNTY State is California, CA.

**4. Solution Strategy.**

The answer for which Cab company XYZ firm should invest can be answered as 2 different ways:

* **First way:** How much money, taxi travels and customers each company made?

1º - Group the variable Company by the variable Profit, and sum the amount of price for each company. This method shows a general perspective about each company in all cities and years.

2º - For taxi travels, it is possible do the same method, however, it is necessary replace the variable Profit by the variable Transaction Id.

3º - For customers, it is possible do the same method, however, it is necessary eliminate the duplications and replace the variable Profit by the variable Customer Id.

* **Second way:** Here it is made a verification of the amount of money and taxi travels by region and season.
* 1º - Group the variable Company by the variable Profit and Region, and sum the amount of price for each company. This method shows a specific perspective about each company in each region. For taxi travels, it is possible do the same method; however, it is necessary replace the variable Profit by the variable Transaction Id.
* 2º - Same method is used than first one, however replace the variable Region by Season. This method shows a specific perspective about each company in each season.

The general and specific overview about price and clients will help the decision about which company XYZ should invest.

Furthermore hypotheses will be created to influence that decision as well.

**5. Top 10 Data Insights.**

**Hypothesis 01:** The number of users of the Yellow Cab plus Pink Cab is the same number of the Total Cab Users at every city.

**False:** There are others Cab companies at each state, which the range of customers in others companies floats between 70% and 99%, depending of the state.

**Hypothesis 02:** The number of users of the Yellow Cab is higher than the company Pink Cab in each city.

**False:** There are cities where the Pink cab company has more customers than the Yellow cab company, cities such: San Diego, Nashville, Sacramento and Pitsburgh.

**Hypothesis 03:** People uses more taxi during New Year's Day than other holidays in all 3 years.

**False:** In total, the holiday which people uses more taxis is Veterans Day, However, splitting this in years, 2016 and 2018, Thanksgiving Day is the holiday which people uses more taxis, and in 2017 the holiday is Veterans Day.

**Hypothesis 04:** The number of users for each company increases WoW (Week over Week).

**False:** The number of users for the company Yellow Cab Company WoW (Week over Week) suffered a decrease of 2.26 percent.

The number of users for the company Pink Cab Company WoW (Week over Week) suffered a decrease of 2.39 percent.

**Hypothesis 05:** The amount of money made by the cab companies increase YoY (Year over Year) and MoM (Month over Month).

**False:** The amount of money made by the company Yellow Cab Company YoY (Year over Year) suffered a decrease of 1.48 percent.

And the amount of money made by the company Pink Cab Company YoY (Year over Year) suffered a decrease of 4.58 percent.

However, the amount of money made by the company Yellow Cab Company MoM (Month over Month) suffered an increase of 1.81 percent.

And the amount of money made by the company Pink Cab Company MoM (Month over Month) suffered an increase of 3.94 percent.

**Hypothesis 06:** The number of travels made by the cab companies increase YoY (Year over Year) and MoM (Month over Month).

**True:** The number of travels made by the company Yellow Cab company YoY (Year over Year) suffered a decrease of 7.06 percent.

And the number of travels made by the company Pink Cab company YoY (Year over Year) suffered a decrease of 8.01 percent.

Also, the number of travels made by the company Yellow Cab Company MoM (Month over Month) suffered an increase of 3.35 percent.

And the number of travels of money made by the company Pink Cab Company MoM (Month over Month) suffered an increase of 4.11 percent.

**Hypothesis 07:** Companies make more profit during second semester of the year.

**True:** In total and in the second semester, the company Yellow Cab make $3770264.7852 (7.64%) more profit than the first semester.

And in the second semester, the company Pink Cab make $2134648.873 (4.33%) more profit than the first semester.

**Hypothesis 08:** People with the amount of income higher than the median uses more taxis.

**False:** People with the amount of income higher than the median uses less taxi than people which has income inferior than median, however this difference is minimal, which is 1 for Pink Cab company and 3 for Yellow Cab company.

**Hypothesis 09:** Men use more cash than women for payment.

**True:** In total, the number of men using cash is 20978 (5.84%) more than women.

**Hypothesis 10:** Young people use more taxis than old people.

**True:** In total, there are 166884 (46.44%) young people more using taxi than old people.

**6. Financial Results.**

Yellow Cab Company has that percentage higher than Pink Cab Company:

* Number of taxi trips (52.86%);
* Profit (78.48%);
* Number of customers (35.73%).

**7. Conclusion.**

In conclusion, Yellow Cab Company has a higher number of taxi trips than Pink Cab Company, as well as a higher profit and higher number of customers.

The data also show that Yellow Cab has superior results in almost all cities, especially in New York, which is the city with the highest profit and higher number of taxi trips.

The same company also stands out from the other company in relation to the season, where the fall or spring (depending which measure is more important) season shows the highest profit and number of taxi trips as well.

Also other insights provided by the hypotheses, inform that Yellow Cab Company present better results than Pink Cab Company, however some of them shows that the number of taxi trips, the profit and the number of customers, for both Companies, are decreasing over the year, giving space for other Cab companies.

Due those all information, it is recommended to invest in Yellow Cab Company.

**8. Next Steps.**

Other project that can be made with this dataset is the exploration data analyses, which identify the best’s attributes in order to apply machine learning algorithms, with the objective to predict the number of taxi trips, the profit or the number of customers for each company.

**3. Project lifecycle along with deadline.**

* Problem understanding
* Data Understanding
* Data Cleaning and Feature engineering
* Model Development
* Model Selection
* Model Evaluation

All those steps are done.