МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ІВАНА ФРАНКА

Факультет прикладної математики та інформатики

ЗВІТ

до індивідуального завдання №3

з дисципліни «Моделі статистичного навчання»

Виконав

студент групи ПМіМ-12:

Зелінський Олександр

Перевірив:

Проф. Заболоцький Т. М.

Львів – 2021

**Хід виконання**

**1.Аналіз даних Weekly**

Розглянемо дані Weekly.

Зображення, що містить текст, будівля

Автоматично згенерований опис

Зображення, що містить текст

Автоматично згенерований опис

Зображення, що містить текст

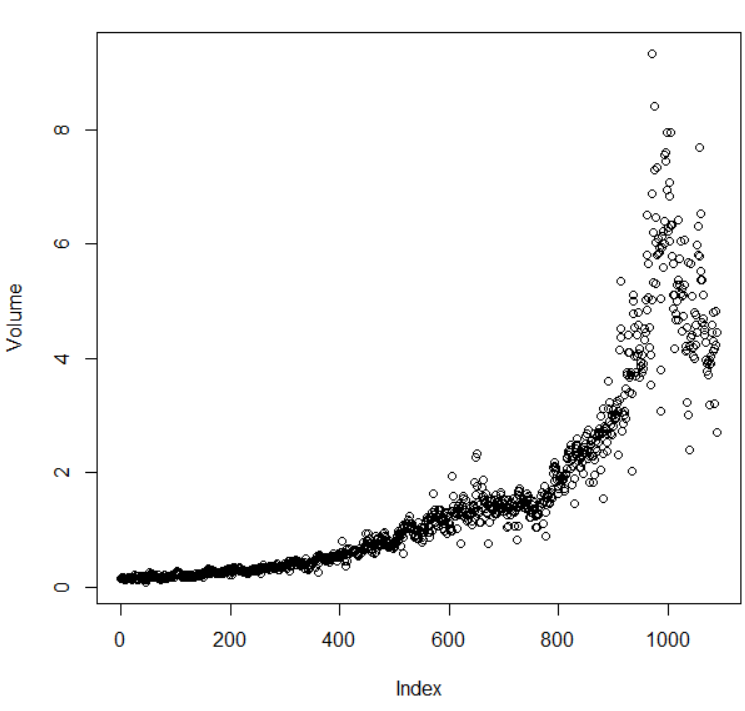
Автоматично згенерований опис

**1.1** Для того, щоб оцінити закономірності розглянемо кореляції між змінними.

**Зображення, що містить текст, газета, документ, знімок екрана

Автоматично згенерований опис**

Чітку кореляцію можна побачити лише між змінними Year та Volume, бо їх значення близькі до 1. З графіка чітко видно не лінійну, а скоріше квадратичну або експонентну залежність.

****

**1.2**

Зображення, що містить текст

Автоматично згенерований опис

Чітко видно, що для заданої моделі найменше p відповідає змінній Lag2, тобто предиктор статистично значущий.

**1.3**

**Зображення, що містить текст

Автоматично згенерований опис**

Після побудови матриці прогнозів та обчислень видно, що загальна частка правильних прогнозів становить 56%, що є не дуже хорошим результатом прогнозування. У ті тижні коли ринок йде вгору, модель правильно прогнозує у 92% випадків, проте у ті тижні коли ринок іде вниз, модель правильно прогнозує у 11.2% випадків.

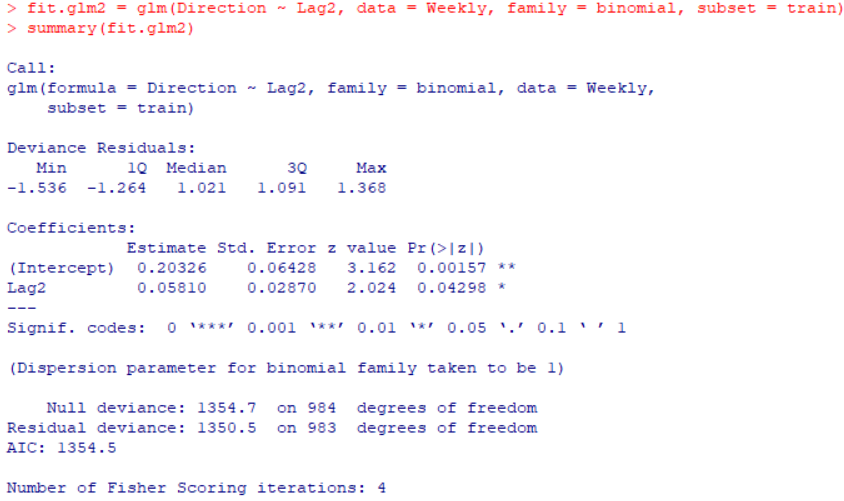
**1.4 Логістична регресія**

Зважаючи на те що дані подано з 1990 року обмеження можна зробити лише на верхню дату тобто 2009 рік.

Зображення, що містить текст

Автоматично згенерований опис

Можемо побачити що тепер тренувальна вибірка складається з 985 записів, а тестова зі 104.



Зображення, що містить текст

Автоматично згенерований опис

Зважаючи на дані з матриці помилок, можна помітити, що модель позначає меншу частину тестових даних як спуск ринку, а більшу як підйом ринку. Загалом частка правильних прогнозів становить 62.5%. У ті тижні коли ринок йде вгору, модель правильно прогнозує у 91.8% випадків, проте у ті тижні коли ринок іде вниз модель правильно прогнозує у 20.9% випадків.

**1.5 Лінійний дискримінантний аналіз**

**Зображення, що містить текст

Автоматично згенерований опис**

Зображення, що містить текст

Автоматично згенерований опис

У цьому випадку видно , що дані є добре розподілені, тобто 55% вибірки позначено як зростання, а 45% як спадання ринку.

Зважаючи на дані з матриці помилок, можна помітити, що модель позначає більшу частину тестових даних як підйом ринку. Загалом частка правильних прогнозів становить 62.5%. У ті тижні коли ринок йде вгору, модель правильно прогнозує у 91.8% випадків, проте у ті тижні коли ринок іде вниз модель правильно прогнозує у 20.9% випадків.

**1.6 Квадратичний дискримінантний аналіз**

**Зображення, що містить текст

Автоматично згенерований опис**

Зображення, що містить текст

Автоматично згенерований опис

Зважаючи на дані з матриці помилок видно, що модель позначає всі тестові дані як зростання ринку. Загалом частка правильних прогнозів становить 58.7%. У ті тижні коли ринок йде вгору, модель правильно прогнозує у 100% випадків, проте у ті тижні коли ринок іде вниз модель не прогнозує правильно взагалі.

**1.7 Класифікатор K-найближчих сусідів з K=1**

Для функції прогнозування формуються дві матриці з предиктора Lag2, які пов’язані окремо з навчальними та окремо тестовими даними.

**Зображення, що містить текст

Автоматично згенерований опис**

Зважаючи на дані з матриці помилок, можна помітити, що частка правильних загалом частка правильних прогнозів становить 50%. У ті тижні коли ринок йде вгору, модель правильно прогнозує у 50.8% випадків, проте у ті тижні коли ринок іде вниз модель правильно прогнозує у 48.8% випадків.

**1.8**

**Зображення, що містить текст

Автоматично згенерований опис**

З наведених вище коефіцієнтів тестових помилок можна зробити висновок, що найбільш відповідними для нашої вибірки виявились моделі лінійного дискримінантного аналізу та логістичної регресії.

**1.9**

**Зображення, що містить текст

Автоматично згенерований опис**

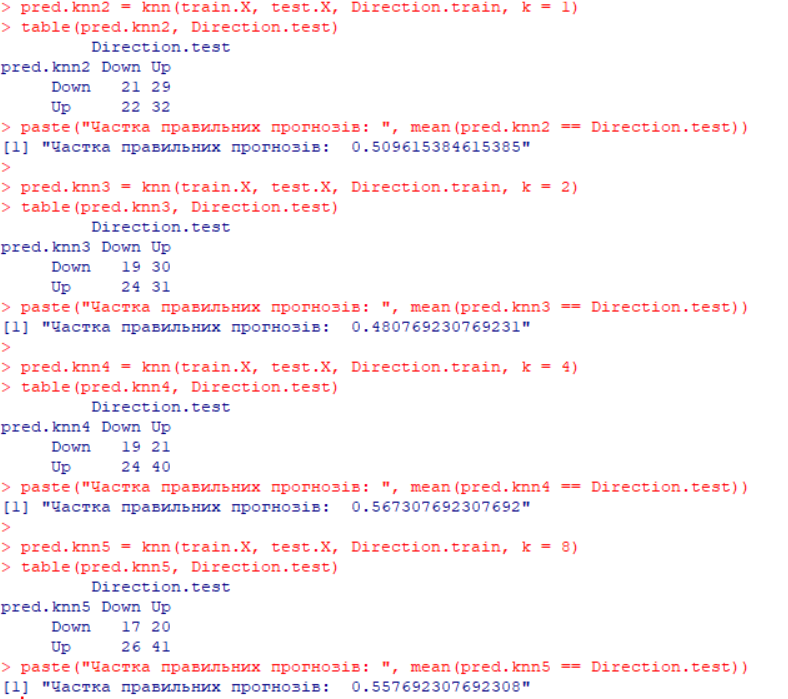
**Зображення, що містить текст

Автоматично згенерований опис**

**Зображення, що містить текст

Автоматично згенерований опис**

Зважаючи на частку правильних прогнозів можна легко помітити, що ці моделі мають гіршу точність ніж початкова модель лінійного дискримінантного аналізу.

****

В результаті виведено матриці помилок при значеннях K= 1, 2, 4, 8. З цих значень добре видно, що для значення K=4 значення частки правильних прогнозів є найкращим, а саме 56.7%.

**2. Модель для передбачення, чи вибране авто має велике або низьке споживання газу на базі даних Auto.**

**2.1** Створено змінну mpg01 та додано її до набору даних Autos.

**Зображення, що містить текст

Автоматично згенерований опис**

Зображення, що містить стіл

Автоматично згенерований опис

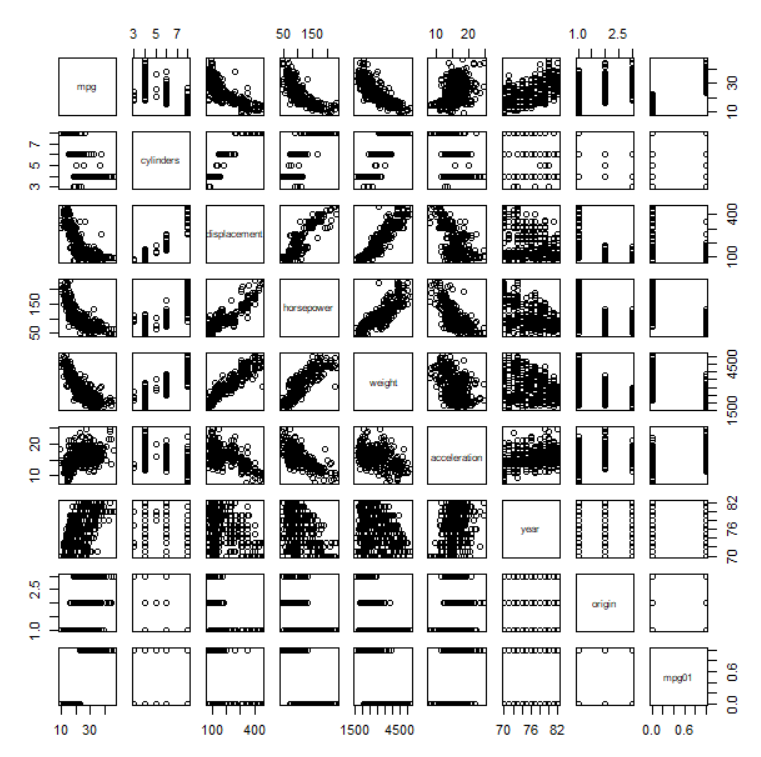
**2.2**

Використано функцію cor(), щоб побачити настільки сильна кореляція між змінними.

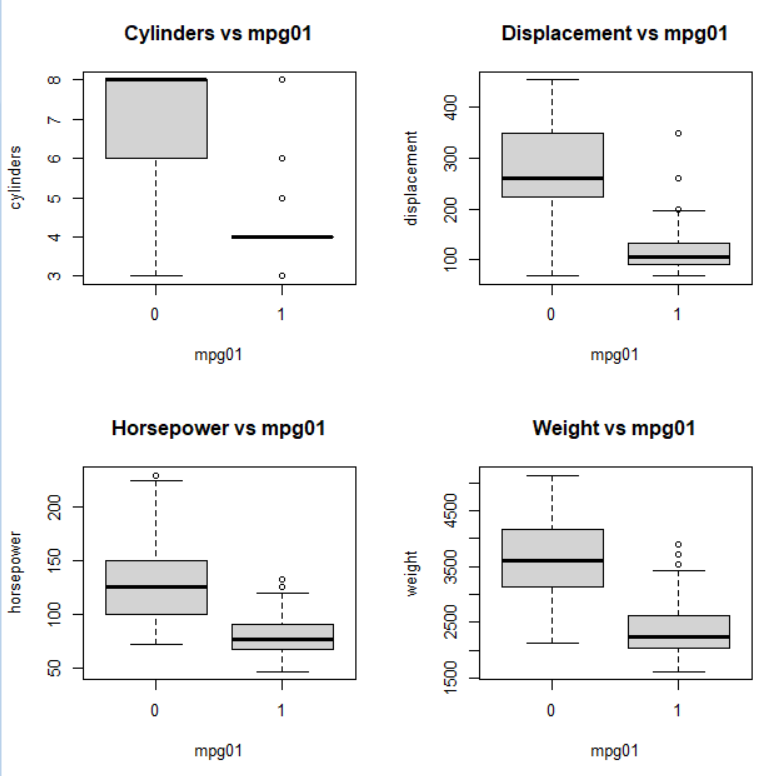
Зображення, що містить стіл

Автоматично згенерований опис

На додачу за допомогою pairs() виведено графічну залежність між всіма змінними.



Отже, з наведених вище даних видно, що існує залежність між mpg01 та cylinders, displacement, horsepower та weight, проте кореляція між ними є від’ємною. Саме тому за допомогою функції boxplot() продовжимо дослідження вищезгаданих змінних.



Всі наведені вище boxplot-и вказують, що змінні cylinders, displacement, horsepower, weight набувають більших значень при значенні змінної mpg01 = 0 ніж при значенні 1. Проте варто зауважити, що для всіх кожної змінної з цього списку існують значення при mpg01 = 1, які є більші за середнє при mpg01 = 0.

**2.3** Вибірку розбито на тестову і навчальну по критерію парний рік чи ні

**Зображення, що містить текст

Автоматично згенерований опис**

**2.4 Лінійний дискримінантний аналіз**

В ролі предикторів було взято змінні cylinders, displacement, horsepower та weight, які є найбільш залежними від mpg01.

**Зображення, що містить текст

Автоматично згенерований опис**

Зображення, що містить текст

Автоматично згенерований опис

Отже в результаті бачимо, що тестова помилка отриманої моделі є 12.6%

**2.5 Квадратичний дискримінантний аналіз**

В ролі предикторів було взято змінні cylinders, displacement, horsepower та weight, які є найбільш залежними від mpg01.

Зображення, що містить текст

Автоматично згенерований опис

Зображення, що містить текст

Автоматично згенерований опис

Отже в результаті бачимо, що тестова помилка отриманої моделі є 13.2%

**2.6 Логістична регресія.**

Зображення, що містить текст

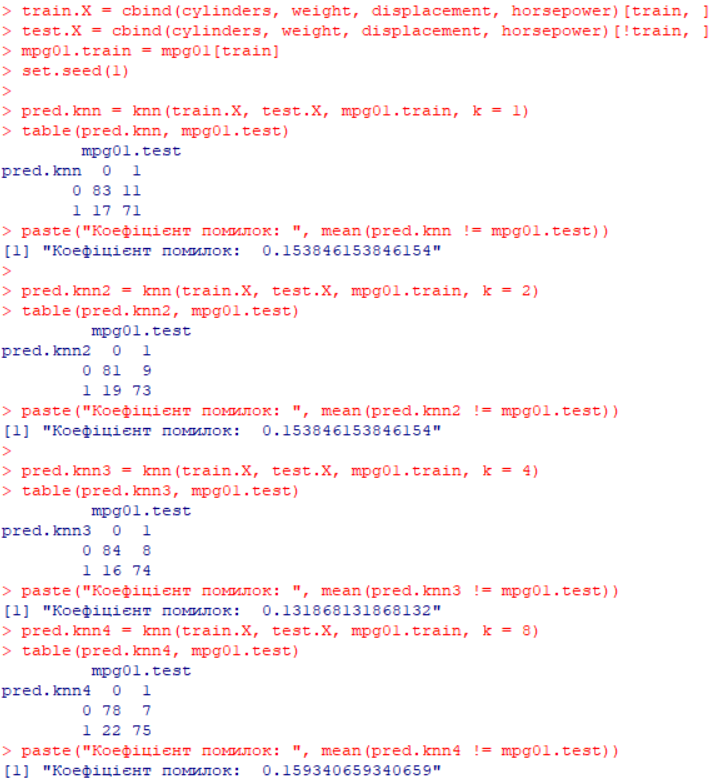
Автоматично згенерований опис

Зображення, що містить текст

Автоматично згенерований опис

Отже в результаті бачимо, що тестова помилка отриманої моделі є 12.1%

**2.7 Метод К-найближчих сусідів з різними значеннями К**



В результаті виведено матриці помилок при значеннях K= 1, 2, 4, 8. З цих значень добре видно, що для значення K=4 значення тестової помилки є найкращим, а саме 13.2%.

**3. Написання функцій**

**3.1**



**3.2**

Зображення, що містить текст

Автоматично згенерований опис

**3.3**

Зображення, що містить текст

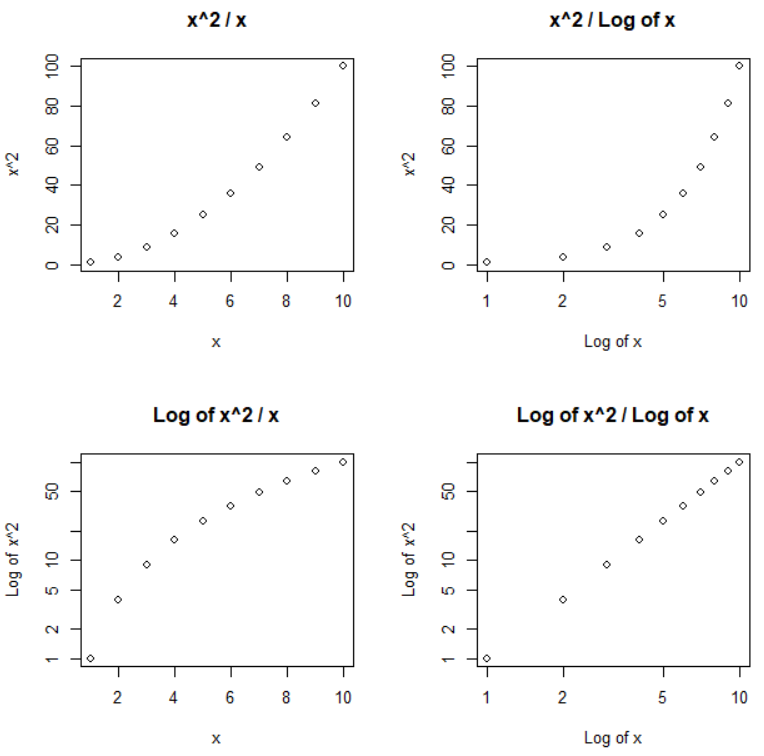
Автоматично згенерований опис

**3.4**

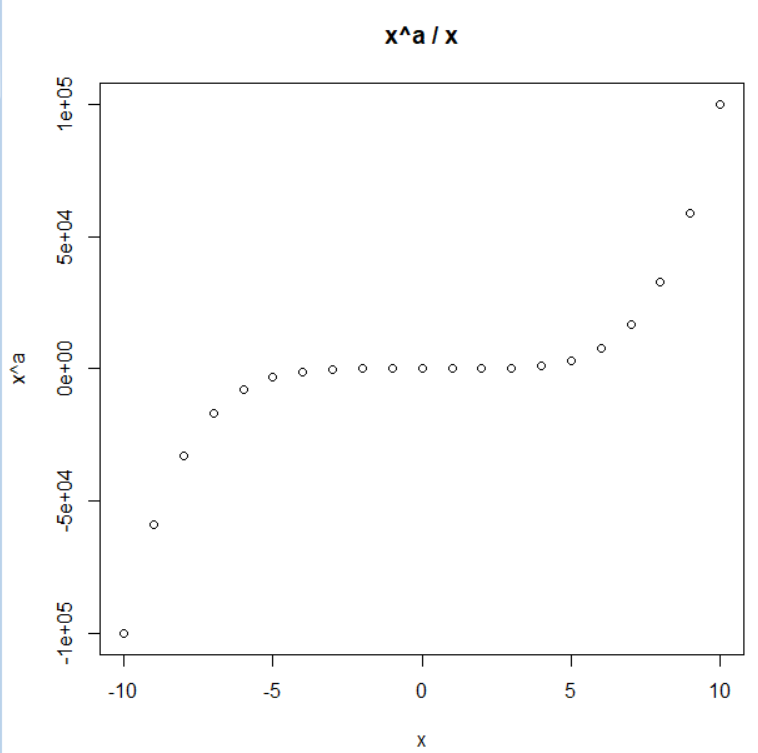
**Зображення, що містить текст

Автоматично згенерований опис**

**3.5**



**3.6** Для прикладу, використано графік де х від -10 до 10 в 5 степені.



**4. Boston**

Зображення, що містить текст

Автоматично згенерований опис

З початку було додано змінну crim01, де її значення 1 якщо crim більше медіани, та 0 якщо менше. Змінну crim01 було додано до Boston. Після цього дані поділені на тренувальну та тестову вибірку.

Для побудови логістичної регресії в ролі предикторів було взято всі змінні, окрім crim01 та crim,

Зображення, що містить текст

Автоматично згенерований опис

З отриманих даних видно, що тестова помилка отриманої моделі є 18.2%. Наступною розглянемо модель лінійного дискримінантного аналізу.

Зображення, що містить текст

Автоматично згенерований опис

Тут тестова помилка буде 13.4%.

Зображення, що містить текст

Автоматично згенерований опис

В цьому випадку тестова помилка 65%. Що означає що дана модель не підходить для даної задачі

Після цього, розглянемо метод К-найближчих сусідів з різними значеннями К

Зображення, що містить текст

Автоматично згенерований опис

Зображення, що містить текст

Автоматично згенерований опис

В результаті виведено матриці помилок при значеннях K= 1, 2, 4, 8. З цих значень добре видно, що для значення K=8 значення тестової помилки є найкращим, а саме 12.6%.