Регрессионая модель

Ядонист Роман

Загрузим данные

Добавим столбец с временем года

1. Поcтроим регрессионную модель, которая позволяет выяснить:

* каким образом потребительская активность зависит от средней заработной платы в регионе, числа активных абонентов беспроводного наземного фиксированного доступа к сети Интернет и уровня безработицы населения.

##   
## Call:  
## lm(formula = Индекс.ПА ~ `Среднемесячная з.п.` +   
## `Число абонентов` + `Уровень безработицы` +   
## Сезон, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -25.5183 -4.1102 -0.0585 4.2416 27.3217   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.874e+01 8.015e-01 73.285 < 2e-16 \*\*\*  
## `Среднемесячная з.п.` -1.019e-05 1.087e-05 -0.937 0.3489   
## `Число абонентов` 5.516e-05 3.053e-05 1.807 0.0711 .   
## `Уровень безработицы` -2.463e-01 5.179e-02 -4.756 2.27e-06 \*\*\*  
## Сезонзима 8.476e+00 6.035e-01 14.046 < 2e-16 \*\*\*  
## Сезонлето 1.324e+01 6.019e-01 21.998 < 2e-16 \*\*\*  
## Сезоносень 1.312e+01 6.018e-01 21.808 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.712 on 989 degrees of freedom  
## Multiple R-squared: 0.4023, Adjusted R-squared: 0.3987   
## F-statistic: 111 on 6 and 989 DF, p-value: < 2.2e-16

* каким образом доля безналичных платежей в торговом обороте зависит от средней заработной платы в регионе, числа активных абонентов беспроводного наземного фиксированного доступа к сети Интернет и уровня безработицы населения.

##   
## Call:  
## lm(formula = Индекс.БП ~ `Среднемесячная з.п.` +   
## `Число абонентов` + `Уровень безработицы` +   
## Сезон, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -28.6934 -3.9745 -0.3483 4.3246 19.7348   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.459e+01 7.853e-01 69.510 < 2e-16 \*\*\*  
## `Среднемесячная з.п.` 1.370e-04 1.065e-05 12.865 < 2e-16 \*\*\*  
## `Число абонентов` 8.419e-05 2.991e-05 2.815 0.004979 \*\*   
## `Уровень безработицы` -1.354e+00 5.075e-02 -26.672 < 2e-16 \*\*\*  
## Сезонзима -1.103e+00 5.913e-01 -1.865 0.062528 .   
## Сезонлето 1.471e-01 5.897e-01 0.249 0.803054   
## Сезоносень 2.135e+00 5.896e-01 3.620 0.000309 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.577 on 989 degrees of freedom  
## Multiple R-squared: 0.5448, Adjusted R-squared: 0.5421   
## F-statistic: 197.3 on 6 and 989 DF, p-value: < 2.2e-16

1. Добавим столбец в датафрейм из 0 и 1. Где за 1 будем считать, что доля онлайн-заявок больше 50%. И он будет показывать склонность населения к онлайн сервисам банка.
2. Добавим столец в датафрейм также из 0 и 1. Где за 1 бедем считать количество ипотечных сделок в регионах больше 500. Этот столбец будет показывать склонность населиня оформлять ипотеку.
3. Построим регрессионную модель, которая позволит выяснить, каким образом склонность населения предпочитать онлайн-услуги банка зависит от средней заработной платы в регионе, числа активных абонентов беспроводного наземного фиксированного доступа к сети Интернет и уровня безработицы населения.

##   
## Call:  
## glm(formula = `Онлайн-заявки dummy` ~ `Среднемесячная з.п.` +   
## `Число абонентов` + `Уровень безработицы` +   
## Сезон, family = "binomial", data = df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.509e+00 3.739e-01 -12.058 < 2e-16 \*\*\*  
## `Среднемесячная з.п.` 5.042e-05 4.639e-06 10.869 < 2e-16 \*\*\*  
## `Число абонентов` 2.576e-05 1.129e-05 2.282 0.0225 \*   
## `Уровень безработицы` 9.314e-02 2.137e-02 4.359 1.31e-05 \*\*\*  
## Сезонзима -1.250e+00 3.046e-01 -4.104 4.07e-05 \*\*\*  
## Сезонлето -1.407e+00 3.127e-01 -4.501 6.75e-06 \*\*\*  
## Сезоносень -6.166e-01 2.688e-01 -2.294 0.0218 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 794.1 on 995 degrees of freedom  
## Residual deviance: 616.1 on 989 degrees of freedom  
## AIC: 630.1  
##   
## Number of Fisher Scoring iterations: 5

## (Intercept) `Среднемесячная з.п.` `Число абонентов`   
## 0.01101149 1.00005042 1.00002576   
## `Уровень безработицы` Сезонзима Сезонлето   
## 1.09761370 0.28655904 0.24477487   
## Сезоносень   
## 0.53977389

1. Построим регрессионную модель, которая позволит выяснить, каким образом склонность населения оформлять ипотеку зависит от средней заработной платы в регионе, числа активных абонентов беспроводного наземного фиксированного доступа к сети Интернет и уровня безработицы населения.

##   
## Call:  
## glm(formula = `Ипотечые сделки dummy` ~ `Среднемесячная з.п.` +   
## `Число абонентов` + `Уровень безработицы` +   
## Сезон, family = "binomial", data = df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.231e+00 3.207e-01 6.957 3.47e-12 \*\*\*  
## `Среднемесячная з.п.` -1.885e-05 3.797e-06 -4.965 6.88e-07 \*\*\*  
## `Число абонентов` 7.656e-05 1.520e-05 5.036 4.74e-07 \*\*\*  
## `Уровень безработицы` -3.111e-01 3.175e-02 -9.799 < 2e-16 \*\*\*  
## Сезонзима 1.895e-01 2.007e-01 0.944 0.345   
## Сезонлето 9.108e-01 2.037e-01 4.471 7.77e-06 \*\*\*  
## Сезоносень 1.820e+00 2.251e-01 8.084 6.25e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1370.3 on 995 degrees of freedom  
## Residual deviance: 1081.0 on 989 degrees of freedom  
## AIC: 1095  
##   
## Number of Fisher Scoring iterations: 5

## (Intercept) `Среднемесячная з.п.` `Число абонентов`   
## 9.3083451 0.9999811 1.0000766   
## `Уровень безработицы` Сезонзима Сезонлето   
## 0.7326140 1.2086700 2.4863133   
## Сезоносень   
## 6.1706827

#### Выводы:

1. В первом задание для предсказания индекса потребеительской активности и индекса безналичных платежей использовали линейную регрессию (так как нужно было предсказать число). Для обоих моделей были использаваны следующие независимые переменные: средняя заработанная плата в регионе, число абонентов доступа к сети интернеи и уровень безработицы населения. Так же построеная модель была зависима от времени года.

В итоге для индекса потребительской активности получили зависимоть от следующих коэффициентов: уровень безработицы населения и от всех ремен года (за исходное время года было взята весна), p-value у этих коэффициентов меньше адекватного уровня значимости. Определили R-squared и он равен 0.4, уровень прогностический силы модели ниже среднего. Зависимости получились следующие:

* при увеличении уровня безработицы на 1, индекс потребительской активности падает на 0.25 при равных прочих;
* при смене времени года: на зиму индекс потребительской активности увеличивается на 8.48 при равных прочих относительно весны; на лето индекс потребительской активности увеличивается на 13.24 при равных прочих относительно весны; на осень индекс потребительской активности увеличивается на 13.12 при равных прочих относительно весны.

Для индекса безналичных платежей получили зависимость от следующих коэффициентов: средняя заработанная плата в регионе, число абонентов доступа к сети интернет, уровень безработицы населения. Так же была определена зависимость изменения от перехода с весны на осень, у остальных времен года (зима и лето) p-value больше адекватного уровня значимости, зависимости от перехода на зиму или лето нет. Определили R-squared = 0.55 уровень прогностический силы модели чуть выше среднего. Зависимости получились следующие:

* при увеличении заработанной платы на 1 рубль, индекс безналичных платежей увеличивается на 1.370e-04 при равных прочих;
* при увеличение числа абонентов на 1, индекс безналичных платежей увеличивается на 8.419e-05 при равных прочих;
* при увеличение уровня безработицы на 1, индекс безналичных платежей падает на 1.35 при равных прочих;
* только при смене на осень, индекс безналичных платежей на 2.16 при равных прочих относительно весны.

1. Для следущего задания необходимо сначала привести целывые переменные к данным состоящими из 1 и 0. В первой модели такая целевая переменная это онлайн-заявки на получения кредита, где 1 означает, что была оформалена онлайн-заявка и 0 соответственно офлайн. Во второй модели целевая переменная будет количество взятых ипотек, где 1 будет означать, что оформлено более 500 ипотек в регионе в месяц и 0 соответственно от 500 и ниже. Для этого выделим начальное количество оформленных ипотек и переведм в количественную шкалу. Далее сравним с количествос в 500 оформленых ипотек. Единица будет показывать регион где активно оформляют ипотки. Построим регрессионные модели для определения какие независимые переменные и насколько влияют на то, что бы человек оформлял кредит онлайн и производил сделки по ипотеке. Соотвественно независимыми переменными в обоих случаях будут: средняя заработанная плата в регионе, число абонентов доступа к сети интернеи и уровень безработицы населения. Так же построеная модель была зависима от времени года.