Preprocessing:

We format the Game Revenue column by removing all missing values and replacing them with zero.

We introduce two new columns:

Restaurants that have installed device

Dates when device is installed

Comparison

KPIs:

* Labor Cost: Presto tries to improve labor efficiency due to increasing labor cost and labor shortages.
* Net Sales: Introduction of the device would help improve the service, hence increase the sales.
* Game Revenue: Introduction of the device would generate Game revenue for Presto

Initial plan was to perform paired t test to compare the performance of the restaurants, before and after the device installation. Since, we have unequal lengths for before and after columns we cannot proceed with this. Even if we interpolate the data to get equal lengths, we would get unstable results.

Comparison 1:

We perform a series of t test for restaurants that have installed device to compare means of groups(when device is installed or not) and evaluate the hypothesis.

|  | **Net Sales** | **Game Revenue (formatted)** | **Labor Cost** | **POS Checks** |
| --- | --- | --- | --- | --- |
| **Dates when device is installed** |  |  |  |  |
| **0** | 0.270285 | 0.000000 | 0.255460 | 0.290073 |
| **1** | 0.244483 | 14.350103 | 0.241053 | 0.251481 |

Game Revenue is obviously greater when device is installed, no need to further investigate.

Net Sales:

* First, we perform F test to compare the variance of the two samples

F test :

Null hypothesis: Variance is equal

Alternate hypothesis: Variances not equal

stats.levene(dfc1['Net Sales'], dfc2['Net Sales'])

LeveneResult(statistic=131.27916254516717, pvalue=2.486799992394135e-30)

At 95% confidence level, p value < 0.05. Thus, reject the null hypothesis and conclude that the variances are not equal.

* We will perform Welch’s t test.

Two tail t test

stats.ttest\_ind(dfc1['Net Sales'], dfc2['Net Sales'], equal\_var=False)

Ttest\_indResult(statistic=14.790184202666891, pvalue=4.7746450014092685e-49)

At 95% confidence level, p value < 0.05. Thus, reject the null hypothesis and conclude that the difference in Net Sales for the two groups is significantly different.

* We suspect the Net Sales to be greater when device is not installed based on mean value.

One tail t test

results = stats.ttest\_ind(dfc1['Net Sales'], dfc2['Net Sales'])

alpha = 0.05

if (results[0] > 0) & (results[1]/2 < alpha):

print ("reject null hypothesis")

else:

print ("do not reject null hypothesis")

At 95% confidence level, p value < 0.05. Thus, reject the null hypothesis and conclude that the Net Sales are greater when device is not installed as compared to when it is.

Labor Cost:

* First, we perform F test to compare the variance of the two samples

F test :

Null hypothesis: Variance is equal

Alternate hypothesis: Variances not equal

stats.levene(dfc1['Labor Cost'], dfc2['Labor Cost'])

LeveneResult(statistic=2.834881593409652, pvalue=0.09224797347736718)

At 95% confidence level, p value > 0.05. Thus, we don’t reject the null hypothesis and conclude that the variances are equal.

* We will perform independent t test.

Two tail t test

stats.ttest\_ind(dfc1['Labor Cost'], dfc2['Labor Cost'])

Ttest\_indResult(statistic=10.230509958784477, pvalue=1.5879025795656944e-24)

At 95% confidence level, p value < 0.05. Thus, we reject the null hypothesis and conclude that the difference in Labor Cost for the two groups is significantly different.

* We suspect the Labor Cost to be greater when device is not installed based on mean value.

One tail t test

results2 = stats.ttest\_ind(dfc1['Labor Cost'], dfc2['Labor Cost'])

alpha = 0.05

if (results2[0] > 0) & (results2[1]/2 < alpha):

print ("reject null hypothesis")

else:

print ("do not reject null hypothesis")

At 95% confidence level, p value < 0.05. Thus, reject the null hypothesis and conclude that the Labor Cost is greater when device is not installed as compared to when it is.

Comparison 2:

We include control groups to our analysis

| **Net Sales** | **Game Revenue (formatted)** | **Labor Cost** | **POS Checks** |
| --- | --- | --- | --- |
| **Dates when device is installed** |  |  |  |  |
| **0** | 3178.305706 | 0.000000 | 642.469401 | 144.878493 |
| **1** | 3050.724421 | 14.350103 | 648.170077 | 136.494738 |

Game Revenue is obviously greater when device is installed, no need to further investigate.

Net Sales:

* First, we perform F test to compare the variance of the two samples

F test :

Null hypothesis: Variance is equal

Alternate hypothesis: Variances not equal

stats.levene(dfc3['Net Sales'], dfc4['Net Sales'])

LeveneResult(statistic=62.312863217650246, pvalue=2.998939018290416e-15)

At 95% confidence level, p value < 0.05. Thus, reject the null hypothesis and conclude that the variances are not equal.

* We will perform Welch’s t test.

Two tail t test

stats.ttest\_ind(dfc3['Net Sales'], dfc4['Net Sales'], equal\_var=False)

Ttest\_indResult(statistic=9.091888714085172, pvalue=1.0148623796500577e-19)

At 95% confidence level, p value < 0.05. Thus, reject the null hypothesis and conclude that the difference in Net Sales for the two groups is significantly different.

* We suspect the Net Sales to be greater when device is not installed based on mean value.

One tail t test

results3 = stats.ttest\_ind(dfc3['Net Sales'], dfc4['Net Sales'], equal\_var=False)

alpha = 0.05

if (results3[0] > 0) & (results3[1]/2 < alpha):

print ("reject null hypothesis")

else:

print ("do not reject null hypothesis")

At 95% confidence level, p value < 0.05. Thus, reject the null hypothesis and conclude that the Net Sales are greater when device is not installed as compared to when it is.

Labor Cost:

* First, we perform F test to compare the variance of the two samples

F test :

Null hypothesis: Variance is equal

Alternate hypothesis: Variances not equal

stats.levene(dfc3['Labor Cost'], dfc4['Labor Cost'])  
  
  
LeveneResult(statistic=66.43277538325681, pvalue=3.717028241390933e-16)

At 95% confidence level, p value < 0.05. Thus, reject the null hypothesis and conclude that the variances are not equal.

* We will perform Welch’s t test.

Two tail t test

stats.ttest\_ind(dfc3['Labor Cost'], dfc4['Labor Cost'], equal\_var=False)

Ttest\_indResult(statistic=-2.215717035696826, pvalue=0.026716101998549513)

At 95% confidence level, p value < 0.05. Thus, we reject the null hypothesis and conclude that the difference in Labor Cost for the two groups is significantly different.

* We suspect the Labor Cost to be lesser when device is not installed based on mean value.

One tail t test

results4 = stats.ttest\_ind(dfc3['Labor Cost'], dfc4['Labor Cost'], equal\_var=False)

alpha = 0.05

if (results4[0] < 0) & (results4[1]/2 < alpha):

print ("reject null hypothesis")

else:

print ("do not reject null hypothesis")

At 95% confidence level, p value < 0.05. Thus, reject the null hypothesis and conclude that the Labor Cost is lesser when device is not installed as compared to when it is.

Machine Learning Model

We apply Multiple Linear regression to capture the relationship between outcome variable (Game Revenue) and predictor variables. The main reason we use this model is because the outcome variable is numeric(continuous).

Preprocessing:

We scale the numeric variables to put them on the same scale so that the variables with larger values do not have a greater effect on the outcome variable.

Diagnosing Multicollinearity:

Multicollinearity occurs when correlation exists between the predictors, this makes our model results invalid. We check VIF for each of the predictors, VIF > 10 indicates multicollinearity.

Net Sales 48.109343

Labor Cost 11.159774

POS Checks 43.165331

Dates when device is installed 1.810061

We remove variable POS checks, this significantly reduces the VIF. The other way to go about it is to perform Principal component Analysis(PCA).

Net Sales 10.305137

Labor Cost 11.055730

Dates when device is installed 1.806626

The VIF for all these variables is close to 10, hence we include all of them in our analysis.

Data Splitting:

We split the dataset into training(70%) and test set(30%).

Initial Model:

sm.OLS(y\_train, X\_train).fit()

|  |  |  |  |
| --- | --- | --- | --- |
| OLS Regression Results | | | |
| **Dep. Variable:** | Game Revenue (formatted) | **R-squared:** | 0.893 |
| **Model:** | OLS | **Adj. R-squared:** | 0.893 |
| **Method:** | Least Squares | **F-statistic:** | 8.346e+04 |
| **Date:** | Wed, 19 Dec 2018 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 14:22:25 | **Log-Likelihood:** | -71523. |
| **No. Observations:** | 30144 | **AIC:** | 1.431e+05 |
| **Df Residuals:** | 30140 | **BIC:** | 1.431e+05 |
| **Df Model:** | 3 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **const** | -4.5983 | 0.042 | -108.797 | 0.000 | -4.681 | -4.515 |
| **Net Sales** | 18.0178 | 0.178 | 101.232 | 0.000 | 17.669 | 18.367 |
| **Labor Cost** | -0.0925 | 0.210 | -0.441 | 0.659 | -0.504 | 0.319 |
| **Dates when device is installed** | 14.5561 | 0.030 | 484.819 | 0.000 | 14.497 | 14.615 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 2093.210 | **Durbin-Watson:** | 1.999 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 9413.538 |
| **Skew:** | 0.183 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 5.713 | **Cond. No.** | 20.6 |

Diagnosing Normality:

For the model, residuals should be normally distributed. We perform Kolmogorov-Smirnov test to check the normality of the residuals.

KstestResult(statistic=0.20644483410750486, pvalue=0.0)

The test is significant as p value < 0.5 at 95%confidencel level, which indicates that residuals are not normally distributed.

Diagnosing Homoscedasticity:

Homoscedasticity means that the errors(residuals) have constant variance. We perform Breusch-Pagan test to check equal variance.

|  |  |
| --- | --- |
| **5235.345279** | Lagrange multiplier statistic |
| **0.000000** | p-value |
| **2111.626240** | f-value |
| **0.000000** | f p-value |

The test is significant as p value < 0.5 at 95%confidencel level, which indicates violation of homoscedasticity ie heteroscedasticity.

The ways to resolve these issues is to perform:

1. Variable transformation(log, box-cox etc.)
2. But, we will perform robust regression that accounts for heteroscedasticity

We include heteroscedasticity consistent covariance matrix(HCCM) so that the model is protected from violation of homogeneity of variance, however the model is not correcting violation of normality.

Model re-run: smf.OLS(y\_train, X\_train).fit(cov\_type='HC0')

|  |  |  |  |
| --- | --- | --- | --- |
| OLS Regression Results | | | |
| **Dep. Variable:** | Game Revenue (formatted) | **R-squared:** | 0.893 |
| **Model:** | OLS | **Adj. R-squared:** | 0.893 |
| **Method:** | Least Squares | **F-statistic:** | 8.108e+04 |
| **Date:** | Wed, 19 Dec 2018 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 14:23:39 | **Log-Likelihood:** | -71523. |
| **No. Observations:** | 30144 | **AIC:** | 1.431e+05 |
| **Df Residuals:** | 30140 | **BIC:** | 1.431e+05 |
| **Df Model:** | 3 |  |  |
| **Covariance Type:** | HC0 |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **const** | -4.5983 | 0.060 | -76.963 | 0.000 | -4.715 | -4.481 |
| **Net Sales** | 18.0178 | 0.283 | 63.745 | 0.000 | 17.464 | 18.572 |
| **Labor Cost** | -0.0925 | 0.242 | -0.382 | 0.702 | -0.567 | 0.382 |
| **Dates when device is installed** | 14.5561 | 0.030 | 491.933 | 0.000 | 14.498 | 14.614 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 2093.210 | **Durbin-Watson:** | 1.999 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 9413.538 |
| **Skew:** | 0.183 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 5.713 | **Cond. No.** | 20.6 |

Interpretation:

* R squared = 0.893

This indicates that 89.3% of variation in Game Revenue is explained by the three predictors.

* Adjusted R squared = 0.893

Adjusted R squared adds penalty to the model if add variables with no or less explanatory power.

The model performs well as we have high value of R squared and the value of R squared is same as Adjusted R squared.

Coefficients:

All the variables are significant except Labor Cost as it has p value < 0.5, this means that Labor cost has coefficient close or equal to zero.

Interpretation of significant coefficients:

* Net Sales = 18.01

This means that if the Net Sales increase by 1 dollar the the Game Revenue will increase by 18.01 dollars.

* Dates when device is installed = 14.55

This means that if when the presto device is installed the Game Revenue is more than 14.55 dollars as compared to when the presto device is not installed.

Insights:

From Comparison:

* For only restaurants that have installed device, the Net Sales and Labor Cost is greater when device is not installed.
* For entire data including control groups, the Net Sales is greater but Labor Cost is less when device is not installed.

From Model:

* The Game Revenue increases with the Net Sales of the restaurant.
* Game Revenue is greater when device is installed as compared to when it is not.