The first model is the baseline performance to compare with the other models. Here, the average price(mean) is used as a predictor. The training data and the dummy regressor fit training means were found to closely match. The calculated average price is 63.81. This seems to indicate a good fit but it would just be an average of the values and it may not be a good predictor.

Next metric will tell us how good the average is as a predictor. The R^2 (coefficient of determination) was used. Results of R^2 of training set of zero and R^2 of the test set as -0.0031, indicating a very poor fit for prediction. Generally, the performance of the test set is expected to be slightly worse than on the training set.

Next metric, will summarize the difference between predicted and actual values using the mean absolute error and mean squared error. MAE (training set)=17.923 and MAE (test set)=19.236. This means MAE on average expected to be off by around $19 if you guessed ticket price based on an average of known values.

MSE (training set)=614.133, MSE (test set)=581.436. Here, we have a slightly better MSE on the test set than we did on the training set.

Next, imputing missing values using median for both training and test sets then scaled each feature to zero mean and unit variance. This was done because of the skew of many of the predictor feature distributions.

The average was used to estimate ticket prices. Result: R^2 of zero for the training and test set. This tells us that the model is over fit, and not a good value for predictive purposes.

Using a linear regression fit model, the median for imputing missing value, explained over 80% of the variance on the training set and over 70% on the test set. The noted lower value for the test set suggests possible overfitting on the training data. We obtained R^2 values for training of 0.817 and for test of 0.720.

A second linear regression fit model using the mean for imputing missing values, yielded R^2 for training and test of 0.817 and 0.716, respectively.

The results of R^2, MAE, MSE for both median and mean do not seem very different, showing it may not make much difference and we may have overtraining in the models. The data suggests trying a subset of features rather than using all of them as inputs.

The linear model needs to be refined because the model was suspected to be overfitting. So, selectKbest was used to select the best features performing f-regression. Results were noted to be worse. selecting a subset of features has an impact on performance. Using the default k of 10 means using 10 features is worse than using all features. Even if different values for k are used, when measuring the performance on the test set and picking the model with the best test set performance, it shows the model is being tuned to the arbitrary test set. This results in a model that works well on the particular quirks of our test set but fails to generalize to new data. This defeats the purpose of keeping the test set of new data to check how well our model might perform on data it hasn’t seen. So, the next step is cross validation to estimate model performance.

Cross validation was performed to build models on k sets of data with k estimates of how the model performs on unseen data but without having to touch the test set.

The results highlight that assessing model performance is inherently open to variability. This will get different results depending on the quirks of which points are in which fold, The advantage of this is that you can also obtain an estimate of the variability, or uncertainty, in the performance estimate.

The pipeline mean cross validation score showed a good value for k is 8. This showed an initial rapid increase with k followed by a slow decline. Also noticeable was the variance of the results greatly increased above k=8. As you increasingly overfit, greater swings in performance were seen as different points move in and out of the training/test folds**.**

Results suggest that vertical drop is the biggest positive feature. Which was consistent with what was seen in the EDA work. Another is the area covered by snow making equipment is a positive feature. The skiable terrain area and trams are negatively associated with ticket price. Does this mean that people will pay less for larger resorts? It could be an effect whereby larger resorts can host more visitors at any one time and so can charge less per ticket. This reminds us about the data on missing information about visitor numbers might still be of interest.

The business needs to be advised whether to undertake further data collection. Gathering data has a cost associated to it so a data quantity assessment tool called learning curve function is used to see how performance varies with differing data set size.

The results from the learning curve function showed that as training set size increases the cross validation score increases. There is an initial rapid improvement in model scores as one would expect, but it’s essentially leveled off by around a sample size of 40-50 showing that there seems to be plenty of data.

Next, the use of random forest regressor model to estimate pipeline performance using cross validation. The final results for the top dominant four features are consistent with the linear model are fast quads, runs, snow making area, vertical drop.

In choosing the model, building a simpler model with fewer features can be advantageous making it easier to sell and explain to stakeholders. We want to build the best linear model and best random forest model, to compare. The final linear regression model performance yielded an MAE of 11.793, with the random forest regression model performance with an MAE of 9.537. With this information the random forest regression model was chosen.

The random forest model has a lower cross validation mean absolute error by almost $1. It also exhibits less variability. Verifying performance on the test set produces performance consistent with the cross validation results. From these results, the random forest regression model was chosen for the business problem to help guide important business decisions.