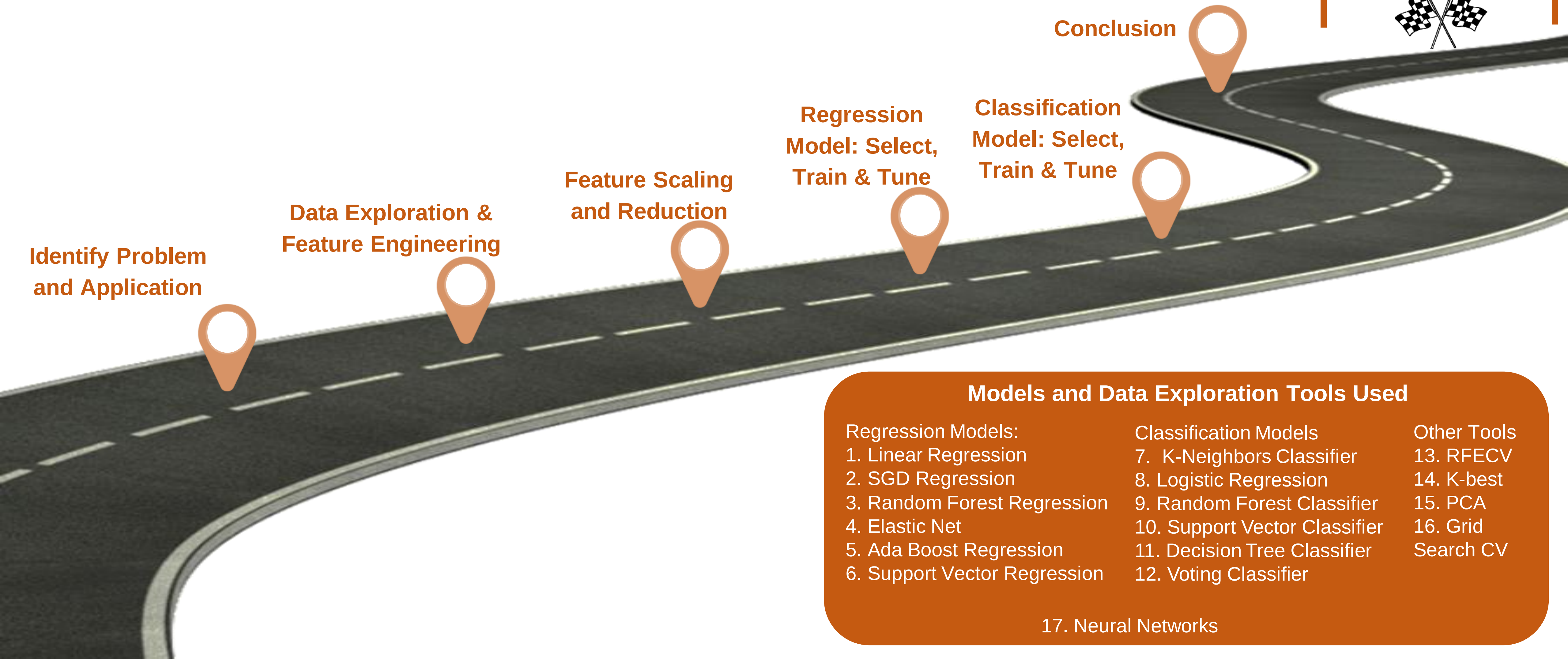
Predicting Pricing and Cancellation for Hotel Reservations

Course Name: SCS 3253-24 Machine Learning
Instructor: Matt MacDonald
Apr 7th 2020

By:
Farrukh Aziz
Harkaran Singh
Sandeep Borkar
Vijaya Chennupati

Pic link: https://i.ytimg.com/vi/Viyp58DMGnq/maxresdefault.jpg

Project Journey



Two Prediction Models for Pricing and Cancellation of Hotel Reservations

Models and Data Exploration Tools Used

Regression Models:
1. Linear Regression
2. SGD Regression
3. Random Forest Regression
4. Elastic Net
5. Ada Boost Regression
6. Support Vector Regression

Classification Models
7. K-Neighbors Classifier
8. Logistic Regression
9. Random Forest Classifier
10. Support Vector Classifier
11. Decision Tree Classifier
12. Voting Classifier

Other Tools
13. RFECV
14. K-best
15. PCA
16. Grid Search CV

17. Neural Networks

Identify Problem and its Application

Important factors while making a hotel reservation

For Customers

- ✓ **Price**
- ✓ Location
- ✓ Reservation Date

For Business

- ✓ **Cancellation**
- ✓ Lead Time
- ✓ Repeat customer

‘Price’ and ‘Cancellation’ are the key factors for any reservation, not just hotel reservation

Problem Identification and its Applicability

Identified the data

- Data Source – Kaggle
- Link to Data Source

https://www.kaggle.com/jessemostipak/hotel-booking-demand

Identified the target variables

- **ADR (Average Daily Rate)**
  - Numerical variable
  - Predict the price of hotel
- **Is Cancelled?**
  - Categorical variable
  - Predict whether a reservation will be cancelled

Application in industry

- ‘Pricing’ and ‘cancellation’ are key parameters for any reservation
- Opportunity to replicate in other industries involving reservations
- Potential to standardize prediction models for Tourism industry

Two Target Variables to Predict: (1) Average Daily Rate (2) Will it be a Cancellation

Data Exploration and Feature Engineering

RAW DATA

119390	31	2	
Number of data points	Number of features	Number of Years	✓ **Industry specific** data related to Hospitality ✓ **'Less-explored'** dataset in Tourism industry vis-a-vis PNR data for Aviation

PROCESSED DATA

Data Clean-Up

1. Blank data
(a) Removed blank 'Country' data (~0.4%)
(b) Replaced blank 'No. Of children' data with median value

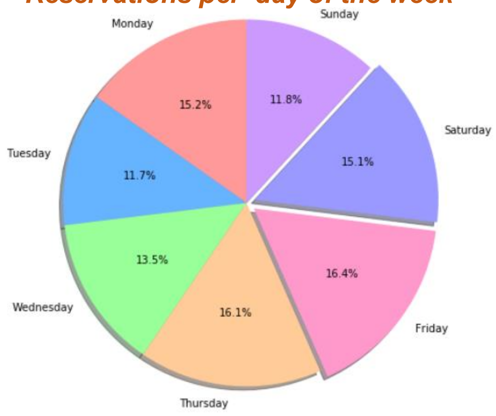
2. Insufficient data: Removed 'Country' data with less than 100 data points

3. Deleted unrelated features

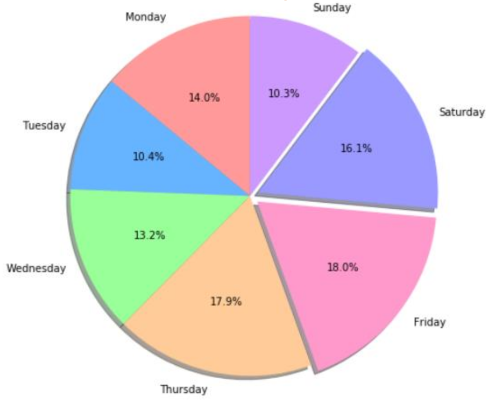
Feature Engineering

1. Generated ***Weekend*** (YES / NO) and ***Day of the week*** from ***Date***

Reservations per 'day of the week'



Cancellations per 'day of the week'

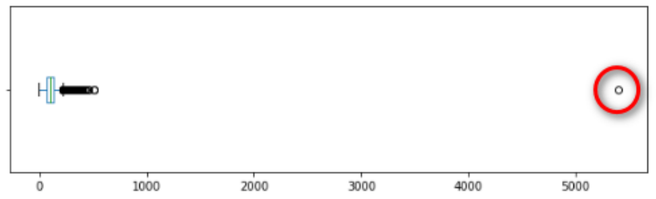


2. Generated ***'Booked through Agent'*** (YES/NO) from ***'Agent ID'***
3. Generated ***'Booked through Company'*** (Y/N) from ***Company ID'***
4. One-Hot Encoding for all categorical features

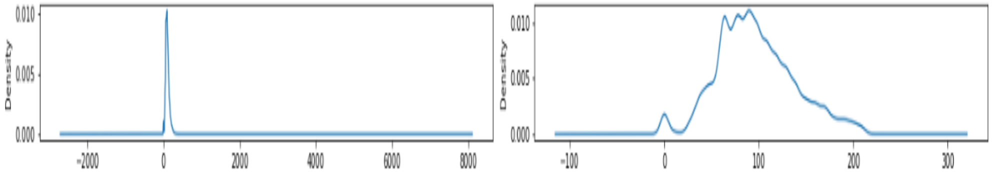
Outlier Removal

Removed all values greater than (Q3 + 1.5 * IQR) or less than (Q1 – 1.5 * IQR)

BOX PLOT: ADR (Average Daily Rate)



BEFORE vs AFTER



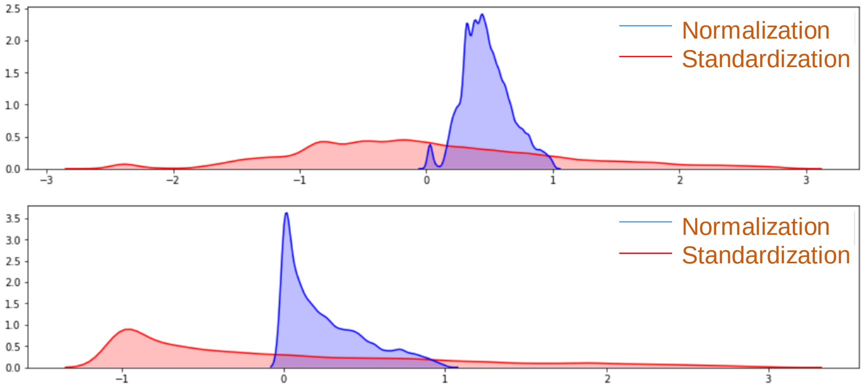
Probability density graphs

Q3 = Third quartile or 75th percentile; Q1 = First quartile or 25th percentile; IQR = Inter-Quartile Range (Q3-Q1)

Identify Problem and Application | **Data Exploration, Feature Engineering** | Feature Scaling and Reduction | Regression Model | Classification Model | Conclusion

Feature Scaling and Reduction

Feature Scaling



Normalization vs Standardization

- ✓ Our dataset has many non-negative features, like price, no. of children etc.
- ✓ Normalization represents non-negative features better
- ✓ Standardization has a much wider spread vis-à-vis normalization

Choose Normalization for feature scaling of our dataset

Feature Reduction / Selection

RFECV

Recursive Feature Elimination with Cross-Validation (RFECV) Output (Regression Only)

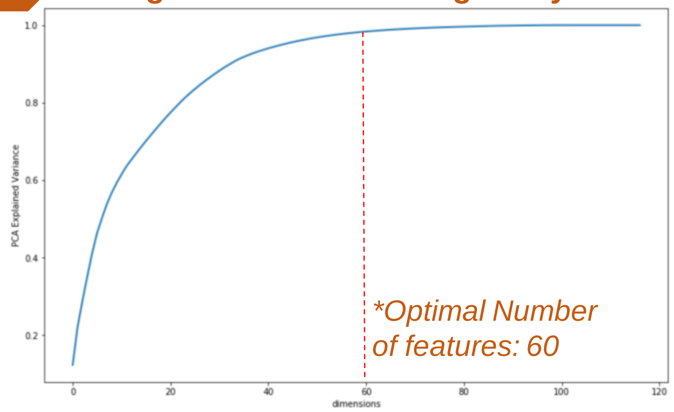
```
Optimal number of features : 80
```

K-BEST

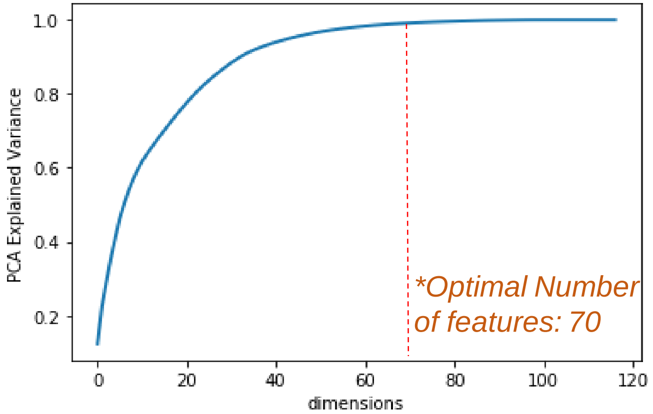
- Regression – Predicting Average Daily Rate: Top 60 features
- Classification – Predicting cancellation: Top 70 features

PCA

Regression: Predict Avg. Daily Rate



Classification: Predict Cancellation



K-best method gave best results for regression and PCA gave best results for classification

Modelling

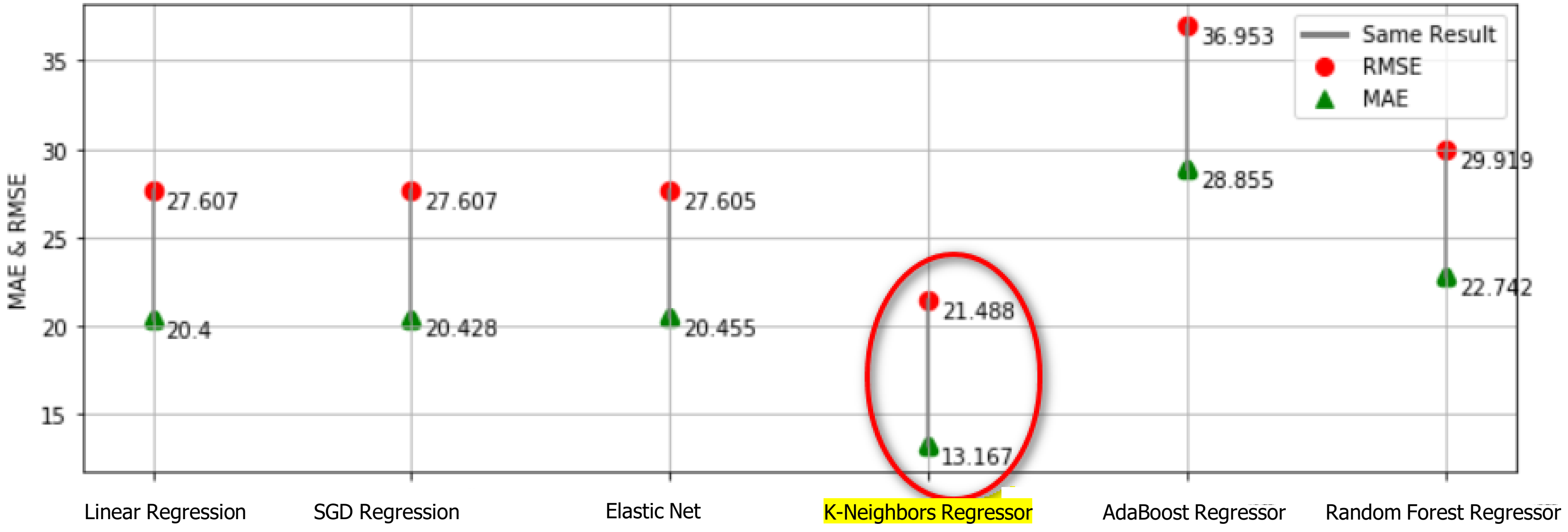
Regression : Predicting Average Daily Date

Select & Train

Identified six models to train :

- Linear Regression
- Stochastic Gradient Descent (SGD) Regression
- Elastic Net
- K-Neighbors Regression
- Ada Boost Regression
- Random Forest Regression

Tune and Compare



Results

K-Neighbors Regression gives best results

- MAE (Mean absolute error): 13.167
- RMSE (Root mean square error): 21.488

Tuned Hyperparameters for K-Neighbors Regression

- leaf_size = 30
- n_neighbors = 3
- p = 2

Modelling

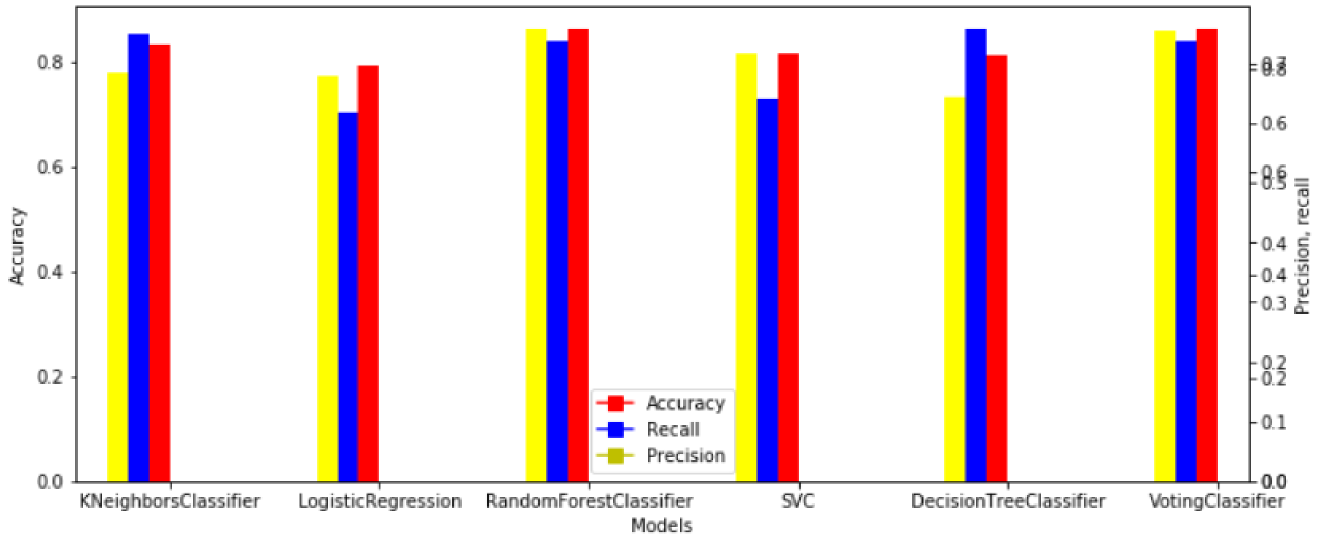
Classification: Predicting Cancellation

Select & Train

Identified six models to train :

- K-Neighbors Classifier
- Logistic Regression
- Random Forest Classifier
- Support Vector Classifier (SVC)
- Decision Tree Classifier
- Voting Classifier

Tune and Compare



	Accuracy Score	Precision	Recall	F1 Score	ROC AUC Score	Model Name
0	0.833628	0.793456	0.748814	0.770489	0.816442	KNeighborsClassifier
1	0.794549	0.786108	0.616979	0.691350	0.758569	LogisticRegression
2	0.864109	0.877509	0.738743	0.802169	0.838706	RandomForestClassifier
3	0.816650	0.828387	0.641199	0.722871	0.781098	SVC
4	0.813515	0.745565	0.758968	0.752207	0.802462	DecisionTreeClassifier
5	0.862588	0.874605	0.737245	0.800072	0.837190	VotingClassifier

Conclusion

Random Forest Classifier (RFC) and **Voting classifier (VC)** give the best results on each of the following parameters

- Accuracy
- Precision
- F1 score
- ROC AUC score

Decision Tree Classifier provides best results for Recall; about 2-3% better than RFC and VC

Results

Classification: Predicting Cancellation

Choose Model

Choose Voting classifier (VC)

- Alongwith Random Forest Classifier (RFC), VC gives the best results for Accuracy and Precision
- The recall score is only ~3% less than the best result
- Choose VC over RFC, since it compares other models (including RFC) and generally provides best results

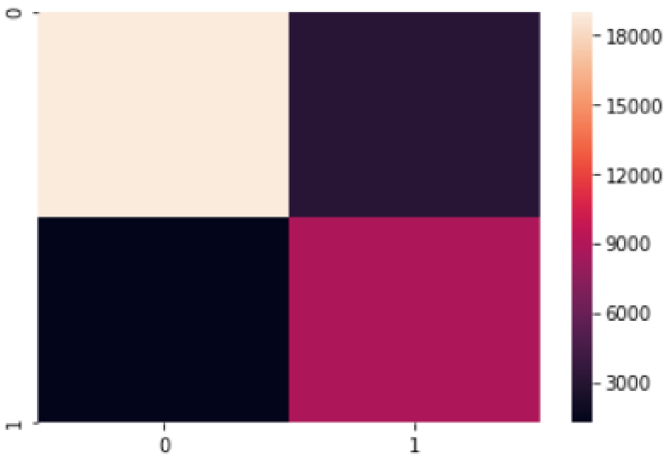
Tuned Hyperparameters

- K-Neighbors Classifier: n_neighbors = 10, weights = ‘distance’
- Logistic Regression: solver = ‘lib linear’, tol = 0.0001, c = 1.0
- Random Forest Classifier: n_estimators = 10, criterion = gini
- Support Vector Classifier: c = 1.0, kernel = rvf, tolerance = 0.001
- Decision Tree Classifier: splitter = ‘best’
- Voting Classifier: voting = ‘hard’

Voting classifier (VC) Results

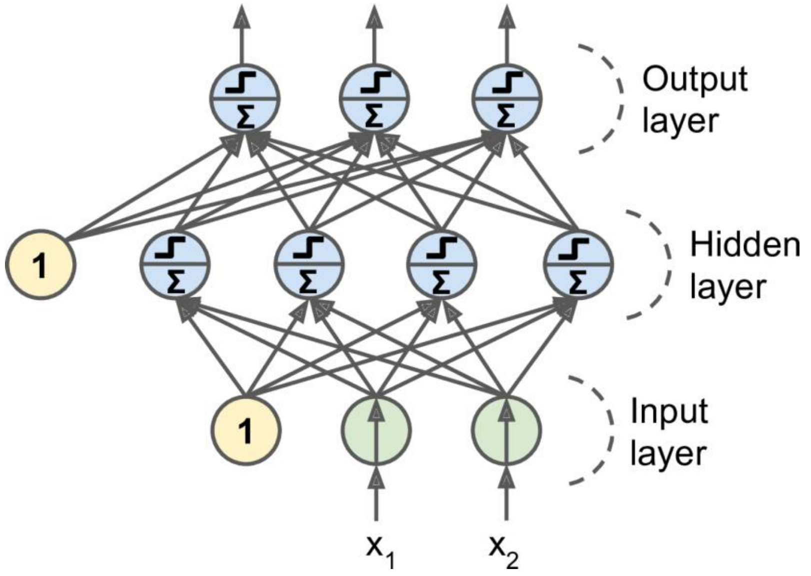
```
accuracy_score : 0.86258807461899
precision_score : 0.874605055292259l
recall_score : 0.7372451102788181
f1_score : 0.800072257598338
roc_auc_score : 0.8371900237068777
```

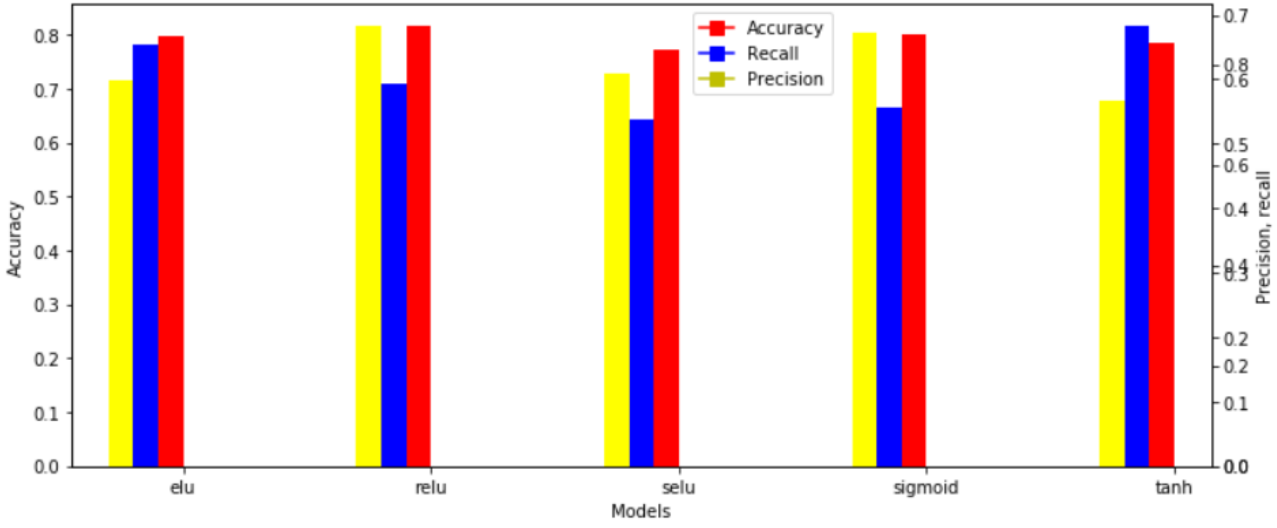
Confusion Matrix



Neural Network (TensorFlow)

Summary





- Neural Networks applied to **predict cancellations** using **three layers** with following perceptron information:
  - Layer 1: 256 units
  - Layer 2: 128 units
  - Layer 3: 1 unit
- Plotted model graph with different activation functions : **elu, relu, selu, sigmoid, tanh**
- All activation scores performed well as recall is over 0.5
- Overall, **'relu'** activation function provided best results followed by **'sigmoid'**

	Accuracy Score	Precision	Recall	F1 Score	ROC AUC Score	Model Name
0	0.797902	0.769264	0.654641	0.697506	0.436833	elu
1	0.816494	0.877258	0.593074	0.695934	0.393021	relu
2	0.772046	0.783504	0.538456	0.626201	0.464521	selu
3	0.800757	0.862087	0.555362	0.663901	0.394712	sigmoid
4	0.785983	0.728180	0.683096	0.695434	0.452558	tanh

Conclusion

Applications of the Model

- As a customer, predict the price of hotel and plan vacation
- As a hotel owner, predict whether a reservation will be cancelled to enable
  - Better Logistics planning
  - Estimation of overbooking required
- Potential to replicate in other reservation related transactions

Model Limitations

- Built on data from two hotels only, which can increase the chances of overfitting
- Method of removing outliers chosen was 1.5 IQR from Q1 and Q3, which could lead to loss of data
- Potential loss of data due to removing ‘Country’ data : blank and less than 100 data points

Challenges and Learnings

- Machine limitation leading to
  - Restriction in using range of hypermeters
  - Unable to use models like Gradient boost and Adaboost
- Virtual working during COVID-19
  - Identified virtual work tools

Future Enhancements

- Apply similar steps for any reservation related model, including flight reservation, appointment reservation etc.
- Can be generalized to any regression and classification problem as our model covers both methods in depth
- Scope for further research to identify underlying clusters in the data

Thank You!