Name: Dikshant Joshi

Lab1 : Causal Forest

Question 1: **Build causal forests. In a table, list the variables you use in the models and explain briefly why you decided to include them (why could it be important w.r.t. treatment-outcome?).**

- For Price as the Outcome Variable

Variables Used	Importance w.r.t Treatment-Outcome
Cleaning fees	➢ Cleaning fees may be related to the complaint line of Airbnb because if a listing is not properly cleaned, guests may be more likely to complain and request a refund. In this case, a high cleaning fee may not necessarily deter guests from booking a listing, but it may increase their expectations for cleanliness and lead to more complaints if those expectations are not met. ➢ Cleaning fees may be related to the nightly price of an Airbnb listing because hosts may set higher prices to offset the cost of cleaning fees. This means that listings with higher cleaning fees may have higher nightly prices, even if they are otherwise comparable to listings with lower cleaning fees.
Instant Bookable (True/False)	➢ Instant bookable listings may be related to the complaint line of Airbnb because they allow guests to book a listing without prior communication with the host. This can result in misaligned expectations between guests and hosts, leading to more complaints. Alternatively, instant bookable listings may be more convenient for guests and result in fewer complaints, depending on the specifics of the listing and guest expectations. ➢ Instant bookable listings may be related to the nightly price of an Airbnb listing because hosts may set higher prices for instant bookable listings, as they may be perceived as more convenient for guests. Alternatively, hosts may set lower prices for instant bookable listings to encourage more bookings, particularly if their listing is less popular or in a less desirable location.
Security Deposit	➢ Security deposits may be related to the complaint line of Airbnb because guests may be more likely to complain if they feel that their security deposit was unfairly withheld or that the listing did not meet their expectations. On the other hand, hosts may be more likely to file a complaint against guests who damage their property, resulting in a security deposit deduction.

Name: Dikshant Joshi

	➢ Security deposits may be related to the nightly price of an Airbnb listing because hosts may set higher prices to offset the cost of the security deposit. Alternatively, hosts may set lower prices to attract more bookings, even if it means forgoing a security deposit.
Extra people (price)	➢ Extra people fee may be related to the complaint line of Airbnb because guests may be more likely to complain if they feel that the fee was unfairly charged or if it was not made clear in the listing description. Alternatively, hosts may complain if guests bring additional people without paying the extra person fee. ➢ Extra people fee may be related to the nightly price of an Airbnb listing because hosts may set higher prices to compensate for the extra person fee. Alternatively, hosts may set lower prices to attract more bookings, even if it means forgoing the extra person fee.
City	➢ City can affect the complaint line of Airbnb because different cities may have different regulations, laws, or cultural norms that impact the guest experience. For example, a city with a high crime rate may lead to more complaints related to safety and security, while a city with strict noise regulations may lead to more complaints related to noise disturbance. ➢ City can affect the nightly price of an Airbnb listing because different cities have different levels of demand, amenities, and attractions. Listings located in popular tourist destinations or business hubs may command higher prices, while listings in less desirable areas may have lower prices.
High Booking(Yes/No)	➢ High booking listings may be more likely to receive complaints simply because they have more guests and activity, increasing the likelihood of a problem occurring. On the other hand, high booking listings may also have more experience and resources to handle complaints and resolve issues quickly, which could lead to fewer overall complaints. ➢ Secondly, high booking listings may have different pricing strategies and expectations compared to low booking listings. Hosts of high booking listings may be more likely to set higher prices due to increased demand and may have higher expectations for guests in terms of behaviour and adherence to house rules.
Treatment	The complaint line binary variable: Coded as 1 for all observations after April, 2016 and 0 otherwise
Price	The outcome variable

Name: Dikshant Joshi

- For Review Score rating as the Outcome variable

Variables Used	Importance w.r.t Treatment-Outcome
Cleaning Fees	➢ High cleaning fee may lead to complaints from guests, as they may feel that the fee is excessive or unfair. In turn, this could result in a negative complaint line for the host. Alternatively, if guests perceive that the cleaning fee is reasonable and that the listing is maintained well, they may be less likely to make a complaint, leading to a more positive complaint line for the host. ➢ Cleaning fees variable can impact review_score_rating in several ways. If guests perceive that the cleaning fee is excessive or unfair, this could lead to a lower review score rating, as they may feel that they did not receive good value for their money. Conversely, if guests perceive that the listing is maintained well and that the cleaning fee is reasonable, this could lead to a higher review score rating, as they may appreciate the cleanliness and attention to detail.
Instant Bookable (True/False)	➢ Regarding treatment, if a host enables the instant bookable feature, this may lead to a higher volume of bookings and potentially a higher likelihood of complaints. This is because guests can book the listing without prior communication or approval from the host, which may result in mismatched expectations or misunderstandings. On the other hand, if a host does not enable the instant bookable feature, they may have more control over the booking process and may be able to better manage guest expectations, potentially leading to fewer complaints. ➢ Regarding outcome, the instant bookable variable can also impact review_score_rating in several ways. If guests appreciate the convenience and ease of using the instant bookable feature, this may lead to a higher review score rating. However, if guests perceive that the listing did not meet their expectations or that the host was unresponsive to their needs, this may lead to a lower review score rating, regardless of whether the listing was instant bookable or not.
City	➢ The city where the Airbnb listing is located can impact the types of complaints that hosts receive. For example, if the listing is in a popular tourist destination, guests may have higher expectations for the cleanliness, amenities, and overall experience. Conversely, if the listing is in a less popular destination, guests may be more forgiving and may have lower expectations. Additionally, the local regulations and laws in the city may impact the types of complaints that hosts receive, such as noise complaints or issues related to parking.

Name: Dikshant Joshi

	➢ The city variable can also impact review_score_rating in several ways. The local culture, cuisine, and attractions in the city may impact the guest's experience and overall satisfaction with the Airbnb listing. Additionally, the quality and availability of public transportation, the safety of the neighborhood, and the proximity to popular tourist destinations may also impact the guest's experience.
High Booking (Yes/No)	➢ High booking variable can impact the types of complaints that hosts receive. If the listing has high booking rates, it may be more challenging for hosts to maintain the property's cleanliness, amenities, and overall condition. This can lead to more complaints from guests about issues such as cleanliness or maintenance problems. Alternatively, if the listing has low booking rates, hosts may be more likely to pay extra attention to maintaining the property's condition to attract more guests. ➢ High booking variable can also impact review_score_rating in several ways. A listing with a high booking rate may be more likely to have a higher review_score_rating if it is popular and meets guest's expectations. However, if the listing has a high booking rate but does not meet guest expectations, it may receive lower review_score_rating due to the high number of guests experiencing issues.
Nightly Price	➢ Nightly price variable can impact the types of complaints that hosts receive. If the listing has a high nightly price, guests may have higher expectations for the cleanliness, amenities, and overall experience. Conversely, if the listing has a low nightly price, guests may be more forgiving and may have lower expectations. Additionally, guests may be more likely to complain about issues related to the value for money if they feel that the nightly price does not reflect the quality of the listing. ➢ Nightly price variable can also impact review_score_rating in several ways. A listing with a high nightly price may be more likely to have a higher review_score_rating if it meets guest's high expectations. However, if the listing has a high nightly price but does not meet guest expectations, it may receive lower review_score_rating due to the guest's perception of poor value for money. On the other hand, a listing with a low nightly price may receive higher review_score_rating if guests feel that they received good value for money.
Host Response Time	➢ Host_response_time variable can impact the types of complaints that hosts receive. If a host has a slow response time, guests may be more likely to complain about issues related to communication, such as difficulty reaching the host or delays in getting responses to their inquiries or complaints. This could also impact the guests' overall experience and perception of the host and listing.

Name: Dikshant Joshi

	On the other hand, if a host has a fast response time, guests may be more likely to feel satisfied with the communication and may be less likely to complain about issues related to communication. ➢ Host_response_time variable can also impact review_score_rating in several ways. A host with a fast response time may receive a higher review_score_rating if guests feel that their communication needs were met, and they were able to get their questions and concerns addressed in a timely manner. Conversely, a host with a slow response time may receive a lower review_score_rating if guests feel that their communication needs were not met, and they were left feeling frustrated or ignored.
Treatment	The complaint line binary variable: Coded as 1 for all observations after April, 2016 and 0 otherwise
Review Score Rating	The outcome variable

Question 2: **What is the causal impact of Airbnb's new complaint line on the ratings?**

The causal impact of Airbnb's new complaint line on the ratings can be visualized by seeing the individual causal prediction column for treated For some rows we can see the effect is positive/negative which says that after introducing new complaint line the ratings got increased/decreased by the values in predictions column.

```{r}
#Question 2: What is the causal impact of Airbnb's new complaint line on the prices?
df2.test%>%
  select(predictions,treatment)%>%
  filter(treatment==1)
```

Description: df [11,821 × 2]

predictions <dbl>	treatment <dbl>
2.77429210	1
0.49460852	1
6.83919856	1
0.86905324	1
1.24301911	1
-0.08427279	1
4.04300682	1
-3.45974889	1
-1.60411594	1
3.10904340	1

Name: Dikshant Joshi

Question 3: **What is the causal impact of Airbnb's new complaint line on the nightly prices?**

The causal impact of Airbnb's new complaint line on the nightly price can be visualized by seeing the individual causal prediction column for treated. For some rows we can see the effect is positive/negative which says that after introducing new complaint line the nightly price got increased/decreased by the $ values in predictions column.

```
#Question 3: What is the causal impact of Airbnb's new complaint line on the prices?
df1.test%>%
  select(predictions,treatment)%>%
  filter(treatment==1)
```

Description: df [12,630 × 2]

predictions <dbl>	treatment <dbl>
-9.37948055	1
-21.66645549	1
3.19768125	1
-59.98989078	1
-16.87521572	1
-17.39209788	1
-25.99519240	1
1.68415042	1
-18.24353095	1
2.00582302	1

1-10 of 12,630 rows Previous 1 2 3 4 5 6 ... 10

Question 4: **For both Question 2 and 3, please estimate the following effects and the standard error of the mean for each estimate:**

a. **The Individual Treatment Effect (ITE)**
b. **The Conditional Average Treatment Effect (CATE)**
c. **The Conditional Average Treatment Effect on Treated (CATET)**
d. **CATE and CATET for New York City, NY**
e. **CATE and CATET for Austin, TX**

- **For Review Rating**
  a. **The Individual Treatment Effect**
     There is no way we can calculate Individual treatment effect because we cannot know the effect of a treated and not treated at the same time.

  b. **The Conditional Average Treatment Effect (CATE)**
     This is difference between average treatment effect on treated and average treatment effect on non-treated which is coming out to be 0.408.

```
# A tibble: 1 x 1
   CATE
  <dbl>
1 0.409
```

```
#Standard error of the mean
std.error(predict(rcf, X2.test))

## predictions
##  0.02303521
```

Name: Dikshant Joshi

c. The Conditional Average Treatment Effect on Treated (CATET)

This is the average of Treatment effect on Treated.

```
#Average Treatment effect on treated
df2.test%>%
  select(predictions,treatment)%>%
  group_by(treatment)%>%
  summarize(TE = mean(predictions))%>%
  summarize(CATET = TE[2])


df2.test%>%
  select(predictions,treatment)%>%
  filter(treatment==1)%>%
  summarize(std_error=std.error(predictions))
```

CATET <dbl>	std_error <dbl>
1.010775	0.02335306

d. CATE & CATET for New York City

- **CATE**

```
df2.test%>%
  select(predictions,treatment,city)%>%
  filter(city=="new-york-city")%>%
  group_by(treatment)%>%
  summarize(TE = mean(predictions))%>%
  summarize(CATE = TE[2]-TE[1])
```

```
# A tibble: 1 x 1
   CATE
  <dbl>                  std_error
1 0.967                1 0.05472127
```

- **CATET**

```
df2.test%>%
  select(predictions,treatment,city)%>%
  filter(city=="new-york-city")%>%
  group_by(treatment)%>%
  summarize(TE = mean(predictions))%>%
  summarize(CATET = TE[2])

df2.test%>%
  select(predictions,treatment,city)%>%
  filter(treatment==1)%>%
  filter(city=="new-york-city")%>%
  summarize(std_error=std.error(predictions))
```

CATET <dbl>	std_error <dbl>
1.436723	0.05723709

Name: Dikshant Joshi

e. CATE & CATET for Austin

- **CATE**

```
df2.test%>%
  select(predictions,treatment,city)%>%
  filter(city=="austin")%>%
  group_by(treatment)%>%
  summarize(TE = mean(predictions))%>%
  summarize(CATE = TE[2]-TE[1])
```

```
# A tibble: 1 x 1
   CATE
  <dbl>
1 0.911
```

```
      std_error
1 0.08890026
```

- **CATET**

```
df2.test%>%
  select(predictions,treatment,city)%>%
  filter(city=="austin")%>%
  group_by(treatment)%>%
  summarize(TE = mean(predictions))%>%
  summarize(CATET = TE[2])

df2.test%>%
  select(predictions,treatment,city)%>%
  filter(treatment==1)%>%
  filter(city=="austin")%>%
  summarize(std_error=std.error(predictions))
```

CATET <dbl>	std_error <dbl>
1.174634	0.08999777

➢ **For Nightly Price**

a. The Individual Treatment Effect

There is no way we can calculate Individual treatment effect because we cannot know the effect of a treated and not treated at the same time

b. The Conditional Average Treatment Effect (CATE)

This is difference between average treatment effect on treated and average treatment effect on non-treated which is coming out to be 0.56.

```
# A tibble: 1 x 1
   CATE
  <dbl>
1 0.563
```

```
#Standard error of the mean
std.error(predict(cf, X.test))

## predictions
##   0.1638863
```

Name: Dikshant Joshi

c. The Conditional Average Treatment Effect on Treated (CATET)

This is the Average of treatment effect on treated.

```
```{r}
#Average Treatment effect on treated
df1.test%>%
  select(predictions,treatment)%>%
  group_by(treatment)%>%
  summarize(TE = mean(predictions))%>%
  summarize(CATET = TE[2])

df1.test%>%
  select(predictions,treatment)%>%
  filter(treatment==1)%>%
  summarize(std_error= std.error(predictions))
```
```

CATET <dbl>	std_error <dbl>
-23.06127	0.1651704

d. CATE & CATET for New York City

- **CATE**

```
df1.test%>%
  select(predictions,treatment,city)%>%
  filter(city=="new-york-city")%>%
  group_by(treatment)%>%
  summarize(TE = mean(predictions))%>%
  summarize(CATE = TE[2]-TE[1])
```

```
# A tibble: 1 x 1
   CATE
  <dbl>
1  3.64
```

```
  std_error
1 0.3600469
```

- **CATET**

```
df1.test%>%
  select(predictions,treatment,city)%>%
  filter(city=="new-york-city")%>%
  group_by(treatment)%>%
  summarize(TE = mean(predictions))%>%
  summarize(CATET = TE[2])

df1.test%>%
  select(predictions,treatment,city)%>%
  filter(treatment==1)%>%
  filter(city=="new-york-city")%>%
  summarize(std_error= std.error(predictions))
```

CATET <dbl>	std_error <dbl>
-20.96141	0.365714

e. CATE & CATET for Austin

- **CATE**

```
df1.test%>%
  select(predictions,treatment,city)%>%
  filter(city=="austin")%>%
  group_by(treatment)%>%
  summarize(TE = sum(predictions))%>%
  summarize(CATE = TE[2]-TE[1])
```

```
# A tibble: 1 x 1
   CATE
  <dbl>
1  15.0
```

```
  std_error
1 0.6946432
```

Name: Dikshant Joshi

- **CATET**

```
df1.test%>%
  select(predictions,treatment,city)%>%
  filter(city=="austin")%>%
  group_by(treatment)%>%
  summarize(TE = sum(predictions))%>%
  summarize(CATET = TE[2])

df1.test%>%
  select(predictions,treatment,city)%>%
  filter(treatment==1)%>%
  filter(city=="austin")%>%
  summarize(std_error= std.error(predictions))
```

CATET <dbl>	std_error <dbl>
-28.68767	0.6960752

Question 5: **In a few paragraphs, describe your thought process in developing the causal forest models, explain, interpret, and discuss your results. What would be your main point if you were to report your results to the Airbnb team? In addition, use bullet points at the end of your write-up to list your key takeaways and lessons learned from the analysis.**

Causal forest models are developed as an extension of random forests to estimate treatment effects. The thought process behind developing these models is to address the challenges faced by traditional methods of causal inference, such as propensity score matching and regression adjustment, in dealing with high-dimensional data and non-linear treatment effects. The causal forest models use a two-step process: first, a random forest is built to estimate the conditional outcome distribution; second, the forest is split into sub-forests based on the heterogeneity of the treatment effects. This approach enables the models to estimate heterogeneous treatment effects across different subgroups.

In the context of Airbnb, causal forest models can be used to estimate the causal impact of Complaint line introduced after 2016 April on metrics nightly price and review score rating. The results from the causal forest models can be interpreted as the conditional average treatment effect (CATE). By examining the CATE, Airbnb can identify the factors that contribute to the effectiveness of Complaint line introduced and if the effect of introducing complaint line had an overall positive or negative impact on the prices and review score rating of the Airbnb.

Main Points:

- The CATE of Complaint line on review rating is 0.4 which is positive. This suggests that introducing a new complaint line on Airbnb lead to higher review ratings, indicating that customers appreciate the effort of Airbnb to address their complaints.
- The CATE of Complaint line on Nightly price is 0.56 which is also positive. This suggests that introducing a new complaint line on Airbnb lead to higher nightly prices, indicating that customers are willing to pay more for a better customer service experience.

Name: Dikshant Joshi

- The CATE of Complaint line on review rating for New York City & Austin is 0.96 & 0.91 respectively which is positive. This suggests that introducing a new complaint line on Airbnb heavily lead to higher review ratings in these cities, indicating that customers appreciate the effort of Airbnb to address their complaints.
- The CATE of Complaint line on Nightly price for New York City & Austin is 3.6 & 14.95 which is also positive. This suggests that introducing a new complaint line on Airbnb lead to higher nightly prices, indicating that customers are willing to pay more for a better customer service experience.
- If we see CATET of Complaint line on review rating and price we can see that the values are 1.01 & -23 respectively which tells us that after introducing complaint line review rating were positively impacted because guests were provided a line to address there complaints and nightly prices of Airbnb got decreased because of more complaints.