Оглавление

Часть 1. Теория и гипотезы.

1.1. Выбор независимых переменных.

1.2. Эффекты взаимодействия и нелинейные эффекты.

Для оценки моделей, которые бы предсказывали вероятность наличия подписки на онлайн кинотеатр были выбраны следующие независимые переменные:

- ***internet*** – доля свободного времени, проводимого в интернете;
- ***age*** – возраст;
- ***TV*** – дамми на факт просмотра телевизора не реже раза в неделю.

Кроме того, были использованы следующие переменные, полученные с помощью нелинейных преобразований:

- ***internet_TV*** – переменная, полученная путем перемножения переменных *internet* и *TV*, принимающая значение 0 в случае, если индивид смотрит телевизор реже одного раза в неделю, и значение, равное доле свободного времени, проводимого в интернете, если индивид смотрит телевизор не реже раза в неделю;
- ***age2*** – переменная, представляющая собой квадрат переменной *age*.

Использование выбранных факторов объясняется их предположительным влиянием на вероятность наличия подписки на онлайн кинотеатр. Так, доля свободного времени, проводимого в интернете, может влиять следующим образом: чем больше доля свободного времени, проводимого в интернете, тем более вероятно, что индивид имеет подписку на онлайн кинотеатр, так как каждый человек при большом количестве свободного времени, проводимом в интернете, потратит его часть на просмотр сериалов или фильмов. Кроме того, большое количество времени, проводимого в интернете, свидетельствует об «интернет-грамотности» индивида, что говорит о том, что, если ему захочется посмотреть фильм или сериал, он сделает выбор именно в пользу просмотра на онлайн платформах, а не по телевизору. Можно также взять во внимание тот факт, что во время «серфинга» по интернету индивид может наткнуться на рекламу, популяризирующую тот или иной онлайн сервис для просмотра фильмов. Более того, индивиду может попасться реклама, которая будет предоставлять особые привилегии при подписке на сервис в виде бонусов, скидок и промокодов, что в значительной степени подогревает желание индивида подписаться. Стоит отметить, что существует такое понятие, как «пиратство контента», но в нашем исследовании мы будем наивно предполагать, что мы имеем дело исключительно с добросовестным потребителем. Итого, предположительно, переменная *internet* будет иметь *положительный* эффект на вероятность наличия подписки на онлайн кинотеатр, то есть $\beta_{internet} > 0$, где $\beta_{internet}$ – коэффициент при переменной *internet*.

В объясняющие факторы включены две переменные, связанные с возрастом, так как предполагается, что вероятность наличия подписки на онлайн кинотеатр будет иметь нелинейную зависимость от возраста. Итак, изначально, когда индивид достаточно молод, вполне вероятно, что у него отсутствует подписка, так как в школьные годы у него нет достаточного количества лишних денег для того, чтобы приобрести себе данную подписку, а в студенческие годы – скорее всего, нет ни лишних денег, ни свободного времени для просмотра сериалов или фильмов. Далее, по мере взросления, заработок индивида растет, доля свободного времени может увеличится, и он может позволить себе подписку на онлайн кинотеатр. При достижении определенного возраста количество свободного времени может достигнуть своего максимума в связи с завершением трудовой деятельности , что может положительно сказаться на вероятность наличия подписки на онлайн кинотеатр, однако реалии данного времени таковы, что ранее упомянутый уровень «интернет-грамотности» у старшего поколения в большинстве своем достаточно низок, и кластер людей определенных возрастов по различным причинам, таким как: отсутствие гаджетов, неумение пользоваться сетью интернет, незнание о существовании сервисов для просмотра фильмом и сериалов ; не могут подключить себе подписку. То есть *гипотеза* относительно влияния возраста на вероятность наличия подписки может быть сформулирована следующим образом: $\beta_{age} > 0$, $\beta_{age2} < 0$, где β_{age} и β_{age2} – коэффициенты при переменных age и age2, соответственно.

Факт просмотра индивидом телевизора не реже раза в неделю, предположительно, косвенно будет влиять на вероятность наличия подписки на онлайн кинотеатр: если индивид смотрит телевизор не реже раза в неделю, то у него, скорее всего, имеется свободное время, которое он предпочитает потратить на просмотр фильмов и сериалов, для чего ему требуется подписка на онлайн сервисы, которые в том числе присутствуют на многих современных телевизорах. Таким образом, предположительно, дамми-переменная *TV* будет иметь *положительный* эффект на вероятность наличия подписки на онлайн кинотеатр, то есть $\beta_{TV} > 0$, *где* β_{TV} – коэффициент при переменной *TV*.

Произведение дамми и доли свободного времени, проводимого в интернете, несёт в себе следующую интерпретацию: в случае, если индивид смотрит телевизор не реже раза в неделю, скорее всего, доля свободного времени, проводимого в интернет, будет в значительной степени больше увеличивать вероятность наличия подписки на онлайн кинотеатр по сравнению со случаем, если индивид смотрит телевизор реже одного раза в неделю. Такое влияние может объясняться тем, что, в случае частого просмотра телевизора, индивид располагает большим количеством свободного времени, а значит и доля свободного времени, проводимого в интернете, наверняка, будет увеличиваться. Кроме того, случай просмотра телевизора индивидом не реже раза в неделю говорит о том, что

он любит смотреть фильмы или сериалы, что увеличивает вероятность наличия подписки. Итого, предположительно, факт просмотра телевизора не реже раза в неделю будет увеличивать эффект влияния переменной *internet*, иными словами, каждая дополнительная единица времени будет сильнее увеличивать вероятность индивида быть подписанным на онлайн кинотеатр в случае частого просмотра телевизора. То есть, *гипотеза* относительно влияния переменной *internet_TV* может быть сформулирована как $\beta_{internet\_TV} > 0$, где $\beta_{internet\_TV}$ - коэффициент прим переменной *internet_TV*.

Таким образом, итоговая модель с зависимой переменной ***sub***, которая равна 1 при наличии у индивида подписки на онлайн кинотеатр и 0 в ином случае, может быть представлена следующим образом:

$$P(SUB_i = 1) = F(\beta_{intercept} + \beta_{internet} * internet_i + \beta_{age} * age_i + \beta_{age2} * age2_i + \\ + \beta_{TV} * TV_i + \beta_{internet\_TV} * internet\_TV_i + \varepsilon_I),$$

где $P(SUB_i = 1)$ – вероятность того, что зависимая переменная равна 1;

$\beta_{intercept}$ – константа, или коэффициент при столбце из единиц;

$\varepsilon_i \sim$ случайные ошибки, независимы;

$F(x_i)$ – соответствующая функция распределения в зависимости от типа выбранной модели, при этом для линейной вероятностной модели $F(x_i) = x_i$.

Часть 2. Линейно-вероятностная модель.

2.1. Оценка модели.

Линейная вероятностная модель представляет собой:

$$P(SUB_i = 1) = \beta_{intercept} + \beta_{internet} * internet_i + \beta_{age} * age_i + \beta_{age2} * age2_i + \\ + \beta_{TV} * TV_i + \beta_{internet\_TV} * internet\_TV_i + \varepsilon_i.$$

Такая модель оценивается *методом наименьших квадратов*, где оцениваемыми параметрами являются коэффициенты $\beta_{intercept}$, $\beta_{internet}$, β_{age}, β_{age2}, β_{TV}, $\beta_{internet\_TV}$.

После оценивания модели линейной вероятности были получены следующие результаты, представленные в *Таблице 1*:

```
Coefficients:
              Estimate  Std. Error t value       Pr(>|t|)
(Intercept)  0.18960458  0.05177772   3.662       0.000253 ***
internet     0.27857973  0.04446439   6.265 0.000000000404 ***
age          0.00450711  0.00169677   2.656       0.007925 **
age2        -0.00002553  0.00001391  -1.836       0.066488 .
TV          -0.16624834  0.02997668  -5.546 0.000000030746 ***
internet_TV -0.06701291  0.05461630  -1.227       0.219889
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4626 on 4994 degrees of freedom
Multiple R-squared:  0.06683,   Adjusted R-squared:  0.06589
F-statistic: 71.53 on 5 and 4994 DF,  p-value: < 0.00000000000000022
```

Таблица 1. Результат оценивания линейной вероятностной модели.

Так как в данной модели нельзя интерпретировать значимость коэффициентов (*Таблица 1*) из-за гетероскедастичности случайно ошибки (подробности в пункте 2.2.), то и нельзя ничего сказать о влиянии факторов на зависимую переменную. Возможные варианты решения данной проблемы для дальнейшей интерпретации указаны ниже впункте 2.2.

2.2. Недостатки и интерпретация модели.

Несмотря на то, что в результате оценивания модели мы получаем несмещенные и состоятельные оценки коэффициентов, линейная вероятностная модель обладает рядом *недостатков*. Во-первых, оценка вероятности может оказаться за пределами значений [0; 1], что делает результаты не интерпретируемыми, так как обычно вероятность должна принимать значения из данного диапазона. Во-вторых, теорема Гаусса-Маркова в данном случае работать не будет, так как ошибка всегда гетероскедастична (разная дисперсия для разных наблюдений). Следовательно, невыполнение условия о теоремы о

гомоскедастичности ошибки приводит к тому, что полученные оценки коэффициентов неэффективные. В-третьих, случайные ошибки в данной модели не могут иметь нормальное распределение, так как оно по своей природе непрерывно, а в линейной вероятностной модели зависимая переменная принимает лишь два значения (1 или 0), следовательно, и случайная ошибка принимает только два значения. Так, нарушение нормальности ошибок приводит к тому, что тестирование гипотез становится невозможным.

Как было сказано выше, полученные оценки несмещенные и состоятельные, однако из-за неадекватности t-статистик (по причине гетероскедастичности случайной ошибки) мы ничего не можем сказать о значимости коэффициентов и, как следствие, интерпретация эффекта влияния переменных через коэффициенты невозможна. По той же причине гетероскедастичности случайной ошибки, F-статистика не будет адекватна, то есть мы не сможем использовать ее для проверки гипотезы об адекватности регрессии. Никакие предпосылки для интерпретации коэффициента детерминации не нарушены, поэтому, теоретически, он может быть проинтерпретирован, однако данная величина не является разумным показателем качества модели, так как в силу специфики данных (бинарная зависимая переменная) нельзя утверждать, какое значение коэффициента считать высоким, а какое – низким. В построенной модели $R^2 = 0.07$, однако это значение ничего не позволит нам сказать о качестве модели.

Вместо перечисленных показателей могут быть использованы альтернативные. Для проверки значимости коэффициентов необходимо проделать следующие шаги: для начала, чтобы побороться с гетероскедастичностью, нужно оценить модель, используя робастные ошибки (или другой альтернативный метод), после чего оценка стандартного отклонения случайной ошибки будет состоятельна. Теперь, беря во внимание факт того, что в наших данных достаточно много наблюдений (5000), мы можем использовать центральную предельную теорему и проверять значимость коэффициентов не через распределение Стьюдента и t-статистику, а через нормальное распределение и z-статистику. Соответственно, в данной модели с робастными ошибками мы можем интерпретировать значимые коэффициенты и влияние подобранных факторов на вероятность наличия подписки на онлайн кинотеатр. Аналогично, используя хи-квадрат распределение и X^2-статистику, мы можем проверить гипотезу об адекватности регрессии. В качестве альтернативы классическому коэффициенту детерминации могут быть использованы другие показатели, например, pseudo-R^2, $R^2_{McFadden}$. Для этого необходимо оценить модель через метод максимального правдоподобия, предварительно сделав предположение о распределении случайных ошибок. Кроме того, можно использовать ROC AUC метод, игнорируя значения оценки вероятности нижу нуля и выше единицы, приравняв их к

пограничным значениям диапазона [0,1].

2.3. Предельные эффекты.

Предельный эффект каждой из используемых переменных представляет собой производную функции по данной переменной. Так, для используемых факторов предельные эффекты будут выглядеть следующим образом:

- *internet*: $ME_{internet,i} = (\partial\, P(\widehat{SUB_i}) = 1)/(\partial\, internet) = \hat{\beta}_{internet} + \hat{\beta}_{internet\_TV} * TV_i$;
- *age*: $ME_{age,i} = (\partial\, P(\widehat{SUB_i}) = 1)/(\partial\, age) = \hat{\beta}_{age} + 2 * \hat{\beta}_{age2} * age_i$;
- *TV*: $ME_{TV,i} = (\partial\, P(\widehat{SUB_i}) = 1)/(\partial\, TV) = \hat{\beta}_{TV} + \hat{\beta}_{internet\_TV} * internet_i$.

1) Первый вариант расчета предельных эффектов осуществляется путем вычисления предельных эффектов переменной для всех наблюдений, а потом взятия среднего из полученных значений, то есть $ME = \frac{1}{n}\sum_{i=1}^{n} ME_i$ (*i=1,…,n; n=5000*, количество наблюдений) для каждой переменной отдельно. Данный вариант исчисления позволяет сказать при каких значениях независимой переменной ее предельный эффект является положительным, а при каких – отрицательным. Для этого возьмем оценки коэффициентов, полученный в пункте 2.1 и подставим их в формулы для индивидуальных предельных эффектов, записанных выше, приравняв ME_i к нулю:

- *internet*: так как в предельном эффекте для переменной *internet* присутствует дискретная переменная, то для нее не имеет смысл описанное вычисление; предельный эффект для *internet* будет положительным для любого значения дамми TV;
- *age*: -(0.00450711) = 2*(-0.00002553) *age_i;

  age_i = 88; так как возраст – неотрицательная величина, то предельный эффект для переменной *age* будет положительным, если возраст индивида менее 88 лет включительно, но он становится отрицательным, если возраст индивида старше 88 лет;
- *TV*: -(-0.16624834) = (-0.06701291) * *internet*;

  $internet_i$ = -2.48084048; так как доля свободного времени, проводимого в интернете, – неотрицательная величина, то предельный эффект для переменной *TV* будет отрицательным для любого разумного значения *internet*;

Рассчитанные предельные эффекты переменных для средних значений представляют собой следующие величины, которые можно видеть в Таблице 3:

AME		
internet	*age*	TV
0.23 $ME > 0 \forall TV$	0.0014 $ME > 0 \forall age \leq 88$	-0.20 $ME < 0 \forall internet$

Таблица 2. Средние предельные эффекты в линейно-вероятностной модели.

В линейно-вероятностной модели нельзя ничего говорить о значимости коэффициентов и, как следствие, предельных эффектов (Таблицы 2, 3), из-за гетероскедастичности случайной ошибки. Так, ничего не зная о значимости, мы не можем ничего сказать и о непосредственном влиянии факторов. Но в задании сказано проинтерпретировать данные средние предельные эффекты. Так при увеличении доли свободного времени, проводимого в интернете, на единицу и увеличении возраста индивида на 1 год вероятность подписки на онлайн кинотеатр увеличивается на 0.24 и 0.0014 соответственно. Что касаемо бинарной независимой переменной, то при прочих равных при условии просмотра телевизора не реже раза в неделю индивид будет иметь подписку на онлайн-кинотеатр с вероятностью на 0.20 меньшей, чем в случае просмотра телевизора реже одного раза в неделю.

2) Второй вариант расчета предельных эффектов осуществляется через средние значения для каждой переменной, то есть в формулы для индивидуальных предельных эффектов подставляются средние по выборке значения фигурирующих там переменных. Для предельного эффекта переменной *internet* входит дамми-переменная TV, поэтому расчет данной характеристики для более внятной интерпретации проводился для двух случаев: для случая, когда индивид смотрит телевизор не реже раза в неделю (*internet.TV)* и для обратного (*internet.noTV)*.

Рассчитанные предельные эффекты переменных для средних значений представляют собой следующие величины, которые можно видеть в Таблице 3:

ME			
internet.TV	*internet.noTV*	*age*	*TV*
0.21	0.28	0.0014	-0.20

Таблица 3. Предельные эффекты среднего в линейно-вероятностной модели.

Так при увеличении доли свободного времени, проводимого в интернете, при условии просмотра индивидом телевизора не реже раза в неделю и при обратном условии вероятность подписки на онлайн кинотеатр увеличивается на 0.21 и 0.28 соответственно. При увеличении возраста индивида на 1 год вероятность наличия подписки увеличивается на 0.0014. Что касаемо бинарной независимой переменной, то при прочих равных при условии просмотра телевизора не реже раза в неделю индивид будет иметь подписку на онлайн-кинотеатр с вероятностью на 0.20 меньшей, чем в случае просмотра телевизора реже одного раза в неделю.

Часть 3. Пробит модель.

3.1. Оценка модели.

Пробит модель представляет собой такую модель, где вероятность зависимой переменной принять значение 1 равна функции распределения для стандартной нормальной величины от переменной индекса:

$$P(SUB_i = 1) = F(\beta_{intercept} + \beta_{internet} * internet_i + \beta_{age} * age_i + \beta_{age2} * age2_i + \\ + \beta_{TV} * TV_i + \beta_{internet\_TV} * internet\_TV_i + \varepsilon_i),$$

где $F(z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-t^2}{2}} dt$ – функция распределения стандартной нормальной случайной величины,

переменная индекса $z_i = \beta_{intercept} + \beta_{internet} * internet_i + \beta_{age} * age_i + \beta_{age2} * age2_i + \\ + \beta_{TV} * TV_i + \beta_{internet\_TV} * internet\_TV_i + \varepsilon_i.$

Функция правдоподобия представляет собой:

$$L = \prod_{i=1}^{n} P(SUB_i = sub_i) = \prod_{i=1}^{n} (F(z_i)^{sub_i} (1 - F(z_i))^{1 - sub_i} \underset{\beta}{\rightarrow} max,$$

$$\beta = \begin{pmatrix} \beta_{intercept} \\ \beta_{internet} \\ \beta_{age} \\ \beta_{age2} \\ \beta_{TV} \\ \beta_{internet\_TV} \end{pmatrix}.$$

Соответственно, такая модель оценивается *<u>методом максимального правдоподобия</u>* относительно коэффициентов β, то есть β – оцениваемые параметры. Полученные оценки будут обладать следующими свойствами:

- состоятельность, $\widehat{\beta_n} \xrightarrow[n \to \infty]{} \beta$;
- асимптотическая несмещенность, $E(\widehat{\beta_n}) \xrightarrow[n \to \infty]{} \beta$;
- асимптотическая эффективность, $Var(\widehat{\beta_n}) = cov(\widehat{\beta_n}) \xrightarrow[n \to \infty]{} I^{-1}(\beta)$;
- асимптотическая нормальность, $\widehat{\beta_n} \xrightarrow[n \to \infty]{} N(\beta; I^{-1}(\beta))$;
- инвариативность относительно гладкого преобразования, $\widehat{g(\beta)} = g(\beta)$.

Результат оценивания пробит модели представлен в *Таблице 4*:

```
Coefficients:
              Estimate  Std. Error z value      Pr(>|z|)
(Intercept) -0.84684898  0.14750337  -5.741 0.00000000940 ***
internet     0.70747811  0.12223608   5.788 0.00000000713 ***
age          0.01322011  0.00484882   2.726        0.0064 **
age2        -0.00007584  0.00003964  -1.913        0.0557 .
TV          -0.50063562  0.08463367  -5.915 0.00000000331 ***
internet_TV -0.07382471  0.15302336  -0.482        0.6295
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6507.3  on 4999  degrees of freedom
Residual deviance: 6170.9  on 4994  degrees of freedom
AIC: 6182.9

Number of Fisher Scoring iterations: 4
```

Таблица 4. Результаты оценивания пробит модели.

Интерпретация результатов оценивания модели (*Таблица 4*) приведена ниже в пункте 3.2.

3.2. Интерпретация результатов.

Интерпретацию результатов стоит начать со значимости коэффициентов модели. Можно увидеть, что значимыми на любом разумном уровне значимости оказались свободный член, коэффициент при доле свободного времени, проводимого в интернете, и коэффициент при дамми на просмотр телевизора не реже раза в неделю. Также, на однопроцентном уровне значимым оказался коэффициент при переменной, отвечающей за возраст индивида. Так, при дальнейшем обсуждении влияния факторов, которое будет вычислено через предельные эффекты, стоит обратить особое внимание на переменную *TV*, отвечающая за факт просмотра телевизора не реже раза в неделю, которая будет иметь отрицательное воздействие на вероятность наличия у индивида подписки на онлайн кинотеатр, так как коэффициент при переменной отрицательный, что противоречит выдвинутой в начале исследования гипотезе. Такое влияние может быть объяснено тем, что индивид, который достаточно часто смотрит телевизор, скорее всего, относится к более старшему поколению, для которого, как мы обсуждали ранее, чуждо понятие онлайн сервисов для просмотра фильмов или сериалов. Влияние же остальных коэффициентов совпало с ранее выдвинутыми гипотезами, а именно: с увеличением доли свободного времени, проводимого в интернете, и возрастом индивида увеличивается вероятность наличия подписки. Стоит отметить, что коэффициент при переменной *age2* незначительно отличается от нуля на уровне значимости 0.05, он в свою очередь имеет отрицательный знак, что свидетельствует о том, что чем старше становится индивид после достижения

определенного возраста, тем меньше вероятность наличия подписки. Коэффициент при переменной internet_*TV* и эффект от относящегося к нему переменной не поддаются интерпретации, так как являются незначимыми, что не позволяет нам на основе данной модели говорить о наличии статистической взаимосвязи. Однако в качестве небольшого дополнительного исследования была построена еще одна пробит модель, аналогичная предыдущей, но без переменной TV. Полные результаты оценивания приведены не будут (необходимый код с комментариями прилагается в R-файле), но важным будет упомянуть, что в такой модели коэффициент при переменной *internet_TV* оказался значимо отличен от нуля при уровне значимости 0.01 и имел отрицательный знак. Такой результат может натолкнуть на мысль о том, что в нашей изначальной модели он был незначим из-за мультиколлинеарности, а более короткая модель без дамми на просмотр телевизора не реже раза в неделю может свидетельствовать о том, что при просмотре телевизора не реже одного раза в неделю, доля свободного времени, проводимого интернете, снижает вероятность подписки на онлайн кинотеатр, и это, кстати, противоречит первоначальной гипотезе о влиянии данной переменной, выдвинутой в начале исследования.

3.3. Оценка вероятности для заданного индивида.

Пусть индивид, для которого будет вычислена вероятность подписки на онлайн кинотеатр, имеет следующие характеристики: доля свободного времени, проводимого в интернете – 0.3, возраст индивида – 23 года, смотрит телевизор не реже раза в неделю. Таким образом, будет рассчитана следующая вероятность: $P(sub_i = 1)$, $individ$: $\{internet = 0.3; age = 23; TV = 1\}$.

Формула, по которой осуществляется расчет имеет вид:

$$P(SUB_{individ} = 1) = F(-0.8468498 + 0.70747811 * internet_{individ} + 0.01322011 * age_{individ} - 0.00007584 * age2_{individ} - 0.50063562 * TV_{individ} - 0.07382471 * internet\_TV_{individ}),$$

где $F(z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-t^2}{2}} dt$

В результате предсказания по пробит модели для индивида с заданными характеристиками мы получили вероятность того, что индивид имеет подписку на онлайн кинотеатр, равную 0.19.

3.4. Средние предельные эффекты.

Оценки предельных эффекты в пробит модели будут иметь разные формулы для непрерывных и дискретных переменных и могут быть вычислены следующим образом:

- для переменной *internet:*

$$AME_{internet} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\, P\left(\widehat{SUB_i}\right)=1}{\partial\, internet} = \frac{1}{n}\sum_{i=1}^{n} f(\hat{z}_i) * \left(\hat{\beta}_{internet} + \hat{\beta}_{internet\_TV} * TV_i\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} f(\hat{z}_i) * (0.70747811 - 0.07382471 * TV_i);$$

- для переменной *age*:

$$AME_{age} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\, P\left(\widehat{SUB_i}\right)=1}{\partial\, age} = \frac{1}{n}\sum_{i=1}^{n} f(\hat{z}_i) * \left(\hat{\beta}_{age} + 2 * \hat{\beta}_{age2} * age_i\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} f(\hat{z}_i) * (0.01322011 - 2 * 0.07382471 * age_i);$$

- для переменной *TV*:

$$\Delta\, AME_{TV} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{P}(SUB_i = 1|TV_i = 1) - \hat{P}(SUB_i = 1|TV_i = 0)\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(f\left(\begin{matrix}-0.84684898 + 0.70747811 * internet_i + 0.01322011 * age_i - \\ -0.00007584 * age2_i - 0.50063562 * TV_i - 0.07382471 * internet_{TV\,i}\end{matrix}\middle| TV_i = 1\right) - \right.$$

$$\left. - f\left(\begin{matrix}-0.84684898 + 0.70747811 * internet_i + 0.01322011 * age_i - \\ -0.00007584 * age2_i\end{matrix}\middle| TV_i = 0\right)\right);$$

где $f(z) = \frac{\partial\, F(z)}{\partial\, t} = \frac{1}{\sqrt{2\pi}}\, e^{\frac{-z^2}{2}}$ – функция плотности для стандартной нормальной случайной величины;

$\hat{z}_i = \beta_{intercept} + \beta_{internet}$*$interneti + \beta_{age}$*$agei + \beta_{age2}$*$age2i + \beta_{TV}$*$TVi + \beta_{internet\_TV}$*$internet\_TVi =$

$= -0.8468498 + 0.70747811$*$internet_i + 0.01322011$*$age_i - 0.00007584$*$age2_i - 0.50063562$ *$TV_i -$

0.07382471*$internet\_TV_i$ – оценка переменной индекса.

Оценки для среднего предельного эффекта представлены в *Таблице 6*:

AME		
internet	*age*	*TV*
0.23	0.0014	-0.19

Таблица 5. Средние предельные эффекты в пробит модели.

Полученные оценки предельных эффектов (*Таблица 6*) могут быть проинтерпретированы следующим образом. Средний предельный эффект показывает, что среднее для всех оценок предельных эффектов по выборке при увеличении возраста

индивида на 1 год и увеличении доли свободного времени на единицу вероятность наличия подписки, увеличивается на 0.014 и на 0.23 соответственно. Что касаемо бинарной независимой переменной, то при прочих равных при условии просмотра телевизора не реже раза в неделю индивид будет иметь подписку на онлайн-кинотеатр с вероятностью на 0.19 меньшей, чем в случае просмотра телевизора реже одного раза в неделю.

3.5. Доля верных предсказаний.

Доля верных предсказаний для нашей модели и для модели линейной вероятности вычисляется как число верных предсказаний, деленное на общее число наблюдений. Предсказание считается верным, если оно совпало с истинным значением зависимой переменной. При этом модель предсказывает не только единицы и нули, а любые числа в данном промежутке. Для того чтобы предсказания тоже были бинарными, мной был взят стандартный уровень cut-off = 0.5, то есть если предсказанное число оказывалось больше данного значения, то оно считалось единицей, если меньше – нулем.

В наивном прогнозе зависимая переменная принимает то значение, которе чаще вссего встречается в наблюдаемой выборке. В данной реализации среднее значение для бинарной зависимой переменной равно 0.36, что ниже порового значения. Соответственно, зависимая переменная примет прогнозную вероятность равную 0.

Так, Таблица 7 показывает долю верных предсказаний для трех типов моделей: пробит, наивной и линейно-вероятностной:

Доля верных предсказаний (%)		
Линейно-вероятностная	*Пробит*	*Наивная*
66.72	66.64	64.46

Таблица 6. Сравнение доли верных предсказаний для трех моделей.

Результаты показывают (Таблица 7), что в линейно-вероятностной модели наибольшая из трех представленных моделей доля верных предсказаний, что может говорить в пользу высокой относительно других проанализированных моделей предсказательной силы линейно-вероятностной модели.

3.6. Проверка гипотезы о значимости предельного эффекта.

Проверим гипотезу о значимости предельного эффекта возраста индивида на вероятность наличия подписки на онлайн кинотеатр на уровне значимости 5 %.

Часть 4. Тестирование корректности спецификации пробит модели.

4.1.Распределение случайных ошибок.

При оценивании модели мы обычно предполагаем, что ошибки распределены нормально. Это позволяет нам интерпретировать статистики для проверки гипотез. В случае же если на самом деле случайные ошибки имеют не нормальное распределение, это может привести к некоторым *негативным последствиям*. В пробит модели оценки коэффициентов выводятся на основе предположения о нормально распределении случайных ошибок, следовательно, при нарушении данного предположения оценки коэффициентов будут несостоятельными.

Проведем тест на нормальность распределения ошибок. В качестве *нулевой гипотезы* выступает гипотеза о нормальности ошибок, а в качестве тестовой статистики используется хи-квадрат распределение.

При проверке гипотезы о нормальном распределении случайных ошибок мы получили p-value = 0.416, что свидетельствует о неотвержении основной гипотезы. Так, мы получили статистические свидетельства в пользу нормальности случайных ошибок в нашей модели.

4.2. Гомоскедастичность случайных ошибок.

Можно предположить, что на дисперсию случайной ошибки будет возраст индивида, то есть при разном разном возрасте будет разный разброс вероятностей наличия подписки. Так, возможно, чем старше индивид, тем больше разброс вероятностей. Такое предположение может быть объяснено следующим образом: когда индивид молодой, у него нет времени и денег для просмотра фильмов и сериалов на онлайн платформе. Когда же индивид взрослеет, у него появляется опция проводить свой досуг за просмотром фильмов и сериалов или же нет. Кроме того, часть той группы, которой нравится поводить свой досуг за просмотром фильмов и сериалов, может заниматься пиратством и не иметь подписки на онлайн кинотеатры, что еще в большей степени увеличивает разборс.

При этом нет оснований утверждать о нелинейной зависимости дохода и стандартного отклонения случайной ошибки, поэтому предполагается следующее: $\sigma_{\varepsilon_i} = age_i * \tau$, где τ – какой-то параметр, τ = const. Так, в тесте на гомоскедастичность случайной ошибки будет проверена следующая гипотеза H_0: $\tau = 0$.

Результаты теста на гомоскедастичность показали, что p-value = 0.012, что свидетельствует о том, что нет оснований отвергать нулевую гипотезу на стандартном уровне значимости 0.01. Так, на основе данного LR теста можно сделать вывод, что, скорее всего, стандартные ошибки гомоскедастичны относительно возраста, то есть разброс вероятностей наличия у индивида подписки на онлайн кинотеатр не зависит от возраста индивида. При этом, если бы после проведения теста были основания отвергать основную гипотезу и ошибки

были гетероскедастичны, то в модели бинарного выбора это могло бы привести к негативным последствиям. Так, в данном случае оценки коэффициентов и, как следствие, предельных эффектов были бы несостоятельными, то есть мы бы не смогли их интерпретировать и делать выводы относительно влияния факторов на исследуемую вероятность. В данном случае гипотеза была проверена с помощью LR теста, однако ему эквивалентен LM тест, который обладает преимуществом над LR тестом. Тест отношения правдоподобия требует оценивания я двух моделей – с ограничениями и без, в то время как для проведения LM теста достаточно оценить лишь модель с ограничениями.

4.3.

4.4. Проверка гипотез о параметрах.

Для проверки гипотез была выбрана переменная age, которая входит в модель как линейно, так и нелинейно – в квадрате.

1) H_0: $\beta_{age} = 0$, или коэффициент при линейной части равен 0. P-value = 0.0069, то есть нет оснований отвергать нулевую гипотезу на стандартном уровне значимости 0.01. Так, результаты теста говорят, что, скорее всего, коэффициент при переменной age статистически значимо не отличается от нуля.

2) $H_0: \begin{cases} \beta_{age} = 0 \\ \beta_{age2} = 0 \end{cases}$, или коэффициенты при линейной и нелинейной частях равняется нулю. P-value = 0.0000011, то есть на любом разумном уровне значимости основная гипотеза отвергается, что говорит о том, что, скорее всего, коэффициенты при переменных age и age2 совместно статистически значимо отличаются от нуля. В целом, возраст индивида статистически значимо влияет на вероятность наличия подписки на онлайн кинотеатр.

3) H_0: $\beta_{age} = k\,\beta_{age2}, k = 2$, или коэффициент при линейной части в два раза больше коэффициента при нелинейной части. P-value = 0.056, следовательно, есть основания отвергать гипотезу на любом разумном уровне значимости. Так, коэффициент при линейной части статистически значимо отличатся от коэффициента при нелинейной части, помноженной на два.

4) $H_0: \begin{cases} \beta_{age} = k\,\beta_{age2}, k = 2, \\ \beta_{TV} = t, t = 0.5. \end{cases}$.

P-value = 0.056, следовательно, есть основания отвергать гипотезу на любом разумном уровне значимости. Так, коэффицент при линейной части статистически значимо отличатся от коэффициента при нелинейной части, помноженной на два, совместно с тем, что коэффициент при дамми статистически значимо отличен от 0.5.

4.5. Проверка гипотезы о совместных моделях для мужчин и женщин.

При помощи LR теста проверим, можно ли оценивать совместную модель для мужчин и женщин.

H_0: коэффициенты в моделях для мужчин и женищн - не различаются.

P-value = 0.046, следовательно, нет основания отвергать гипотезу на уровне значимости 0.01. Так, коэффициенты в моделях для мужчин и женщин не различаются.

4.6. Проверка гипотезы о совместных моделях людей, проживающих в разных населённых пунктах.*

При помощи LR теста проверим, можно ли оценивать совместную модель для людей, проживающих в разных населенных пунктах.

1) Сравним модели для людей, которые приживают в Village и City.

H_0: коэффициенты в моделях для людей, проживающих в Village и City - не различаются.

P-value = 0, следовательно, нет основания отвергать гипотезу на любом уровне значимости. Так, коэффициенты в моделях для людей, проживающих в Village и City, не различаются.

2) Сравним модели для людей, которые приживают в Village и Captital.

H_0: коэффициенты в моделях для людей, проживающих в Village и Capital - не различаются.

P-value = 0, следовательно, нет основания отвергать гипотезу на любом уровне значимости. Так, коэффициенты в моделях для людей, проживающих в Village и Capital, не различаются

3) Сравним модели для людей, которые приживают в City и Captital.

H_0: коэффициенты в моделях для людей, проживающих в City и Capital - не различаются.

P-value = 0, следовательно, нет основания отвергать гипотезу на любом уровне значимости. Так, коэффициенты в моделях для людей, проживающих в City и Capital, не различаются.

Таким образом, коэффициенты в моделях для людей, проживающих в разных населенных пунктах, не различаются.

Часть 5. Логит модель

5.1. Оценка модели.

Логит-модель представляет собой такую модель, где вероятность зависимой переменной принять значение 1 равна логистической функции от переменной индекса:

$$P(SUB_i = 1) = F(\beta_{intercept} + \beta_{internet} * internet_i + \beta_{age} * age_i + \beta_{age2} * age2_i + \\ + \beta_{TV} * TV_i + \beta_{internet\_TV} * internet\_TV_i + \varepsilon_i),$$

где $F(z) = \frac{e^z}{1+e^z}$,
переменная индекса $z_i = \beta_{intercept} + \beta_{internet} * internet_i + \beta_{age} * age_i + \beta_{age2} * age2_i +$
$+ \beta_{TV} * TV_i + \beta_{internet\_TV} * internet\_TV_i + \varepsilon_i.$

Таким образом, главное <u>*отличие*</u> логит модели от пробит модели состоит в используемой функции распределения в качестве функции от переменной линейного индекса.

Функция правдоподобия представляет собой:

$$L = \prod_{i=1}^{n} P(SUB_i = sub_i) = \prod_{i=1}^{n} (F(z_i)^{sub_i}(1 - F(z_i))^{1-sub_i} \underset{\beta}{\rightarrow} max,$$

$$\beta = \begin{pmatrix} \beta_{intercept} \\ \beta_{internet} \\ \beta_{age} \\ \beta_{age2} \\ \beta_{TV} \\ \beta_{internet\_TV} \end{pmatrix}.$$

Результат оценивания пробит модели представлен в *Таблице 7*:

```
Coefficients:
              Estimate  Std. Error z value      Pr(>|z|)
(Intercept) -1.36629353  0.24326529  -5.616 0.00000001949 ***
internet     1.13279037  0.19716874   5.745 0.00000000918 ***
age          0.02132131  0.00800693   2.663       0.00775 **
I(age^2)    -0.00012070  0.00006537  -1.846       0.06483 .
TV          -0.83892994  0.13918355  -6.028 0.00000000167 ***
internet:TV -0.07420132  0.25026699  -0.296       0.76686
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6507.3  on 4999  degrees of freedom
Residual deviance: 6170.8  on 4994  degrees of freedom
AIC: 6182.8

Number of Fisher Scoring iterations: 4
```

<u>Таблица 7</u>. Результаты оценивания логит модели.

В логит модели нельзя напрямую интерпретировать коэффициенты как влияние факторов на зависимую переменную. Однако, для начала, мы можем посмотреть (Таблица 7) на значимость и увидеть, что статистически значимо отличными от нуля на любом уровне значимости оказались коэффициенты при свободном члене, при доле свободного времени, проводимого в интернете, при дамми на просмотр телевизора не реже раза в неделю. На уровне значимости 0.01 оказался статистически значим коэффициент при переменной, отражающей возраст индивида. Так же, коэффициент при переменной age2 оказался значим на уровне значимости 0.05. Так, для них мы сможем посмотреть на знак коэффициентов и сказать следующее: чем старше индивид и чем больше свободного времени он проводит в интернете, тем, скорее всего, более вероятно он будет подписан на онлайн кинотеатр, так как знак при этих коэффициентах - отрицательный. Просмотр телевизора не реже раза в неделю имеет негативный эффект: так как знак при коэффициенте отрицательный. Так же при достижении определенного возраста с каждым последующим годом вероятность подписки индивида на онлайн кинотеатр снижается. О влиянии других факторов говорить будет неуместно, так как коэффициенты при них незначимы.

5.2. Отношение шансов.

Для вычисления оценки изменения отношения шансов по независимой переменной входящей линейно из модели был убран квадрат возраста индивида, чтобы посчитать необходимое изменение по входящему теперь только линейно возрасту индивида.

Оценка изменения отношения шансов по возрасту индивида оказалась равна 1.006792. Таким образом, изменении дохода мужа на 1 год отношение шансов, при прочих равных, изменится в 1.008 раз, то есть увеличится. Иными словами, отношение вероятности наличия подписки к вероятности того, что индивид не подписан на онлайн кинотеатр, увеличится. Также это можно интерпретировать как рост вероятности подписки и/или увеличение вероятности наличия подписки при росте возраста индивида на единицу.

Часть 6. Система бинарных уравнений.

6.1. Оценка системы уравнений.

Система бинарных уравнений будет состоять из двух пробит-моделей:

$$P(SUB_i = 1) = F(\beta_{intercept1} + \beta_{internet} * internet_i + \beta_{age} * age_i + \beta_{age2} * age2_i + + \beta_{TV} * TV_i + + \beta_{internet\_TV} * internet\_TV_i + \varepsilon_i),$$

$$P(TV_i = 1) = F(\beta_{intercept2} + \beta_{internet} * internet_i + \beta_{age} * age_i + \beta_{age2} * age2_i + \beta_{residence} * residence_i + \varepsilon_i).$$

где F(z) – функция распределения стандартной нормальной случайной величины.

Первое уравнение системы аналогично анализируемому ранее: в качестве зависимой переменной выступает факт подписки на онлайн кинотеатр, а в качестве влияющих факторов – доля свободного времени, проводимого в интернете, возраст и квадрат возраста индивида, факт просмотра телевизора не менее раза в неделю и переменная взаимодействия доли свободного времени, проводимого в интернете и дамми просмотр телевизора не менее раза в неделю. Во втором уравнении системы объясняемой переменной является факт просмотра телевизора не менее раза в неделю, а объясняющими – переменная, отвечающая за тип местности, где проживает индивид (предположительно, вероятность просмотра телевизора не реже раза в неделю увеличивается от индивида, проживающего в столице, к индивиду, проживающему в городе, от индивида, проживающего в городе, к индивиду, проживающему в деревне, так как вероятность наличия интернета в деревне намного выше, чем в городах; плюс люди, проживающее столицах и мегаполисах в большинстве очень заняты, а так же предпочитают интернет просмотру телевизора), возраст (предположительно, молодое поколение не так сильно заинтересованы в просмотре телевизора, как старшее поколение) и доля свободного времени, проводимого в интернете (предположительно, большая доля времени, проводимого в интернете, снижает вероятность просмотра телевидения). Результаты оценивания системы бинарных уравнений представлены в Таблице 8:

```
EQUATION 1
Link function for mu.1: probit
Formula: sub ~ internet + age + I(age^2) + TV + internet:TV

Parametric coefficients:
               Estimate  Std. Error z value            Pr(>|z|)
(Intercept) -1.19444071  0.12438121  -9.603 <0.0000000000000002 ***
internet     1.22919078  0.08514745  14.436 <0.0000000000000002 ***
age         -0.00298406  0.00424285  -0.703               0.482
I(age^2)    -0.00003758  0.00003488  -1.077               0.281
TV           1.10929755  0.04733834  23.433 <0.0000000000000002 ***
internet:TV -0.04741403  0.08580855  -0.553               0.581
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
EQUATION 2
Link function for mu.2: probit
Formula: TV ~ internet + age + I(age^2) + residence

Parametric coefficients:
                   Estimate  Std. Error z value            Pr(>|z|)
(Intercept)       0.23665066  0.13428770   1.762              0.0780 .
internet         -1.23943580  0.07252069 -17.091 <0.0000000000000002 ***
age               0.01188904  0.00477429   2.490              0.0128 *
I(age^2)          0.00002194  0.00003967   0.553              0.5802
residenceCity    -0.06478036  0.02854734  -2.269              0.0233 *
residenceVillage -0.29278566  0.03420327  -8.560 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


n = 5000  theta = -0.988(-0.998,-0.93)  tau = -0.9(-0.956,-0.761)
total edf = 13
```

Таблица 8. Результаты оценивания системы уравнений.

Интерпретация результатов оценивания системы бинарных уравнений (Таблица 8) приведена ниже в пункте 6.2.

6.2.Интерпретация результатов.

Для начала обратимся к первому уравнению системы, в котором значимо отличными от нуля оказались коэффициент при дамми на просмотр телевизора не реже раза в неделю, коэффициент при свободном члене и коэффициент при доли свободного времени, проводимого в интернете, на любом разумном уровне значимости. Коэффициенты при других переменных статистически не значимы, что не дает нам статистических свидетельств в пользу влияния данных факторов на вероятность наличия подписки на онлайн кинотеатр. Так как по коэффициентам пробит модели мы можем сделать вывод лишь о направлении влияния

факторов, но не его силе, посмотрим на знаки коэффициентов, которые статистически значимо отличаются от нуля. Так, коэффициент при доле свободного времени, проводимого в интернете, положительный, что может свидетельствовать о том, что, скорее всего, при прочих равных, чем больше свободного времени индивид проводит в интернете, тем выше вероятность подписки на онлайн кинотеатр. Дамми-переменная на просмотр телевизора не реже раза в неделю имеет положительный коэффициент, что может свидетельствовать о том, что, при прочих равных, для индивида, который смотрит телевизор не реже раза в неделю, скорее всего, вероятность работать выше, чем для индивида, который смотрит телевизор реже.

В результате оценивания второго уравнения системы мы получили четыре статистически значимо отличных от нуля коэффициента – при доли свободного времени, проводимого в интернете, на любом разумном уровне, при возрасте индивида на уровне значимости 0.01, при коэффициенте для индивидов, проживающих в городе, при коэффициенте для индивидов, проживающих в деревне. Остальные же коэффициенты оказались статистически незначимы, что не дает нам интерпретировать результаты. Интерпретация знака коэффициента при доли свободного времени, проводимого в интернете, может говорить о том, что, при прочих равных, индивид, который проводит много свободного времени в интернете, скорее всего, будет иметь подписку на онлайн кинотеатр с меньшей вероятностью, чем индивид, который не любит проводить свое время в интернете, так как коэффициент при переменной отрицательный. Положительный коэффициент при возрасте индивида может говорить о том, что, скорее всего, вероятность подписки на онлайн кинотеатр увеличивается с возрастом индивида. Отрицательный коэффициент при переменной для индивидов, проживающих в городах и в деревнях, говорит о том, что люди, живущие в данной местности, имеют подписку на онлайн кинотеатр с меньшей вероятностью.

Оценка коэффициента корреляции между случайными ошибками рассматриваемых уравнений представлена в виде theta = -0.988, что может свидетельствовать о том, что уравнения связаны.

6.3. Оценка вероятности для заданного индивида.

В качестве индивида для расчета оценки вероятности был выбран индивид 23 лет (age = 23) с долей свободного времени, проводимым в интернете, равным 0.3 (internet = 0.3) и без просмотра телевизора не реже раза в неделю (TV = 0), который проживает в деревне (residence = “Village”).

1) Оценка вероятности подписки на онлайн кинотеатр оказалась равна 0.18.

2) Оценка вероятности того, что индивид смотрит телевизор не реже раза в неделю оказалась равна 0.44.

3) * Оценка вероятности того, что индивид и имеет подписку, и смотрит телевизор не реже раза в неделю равна 0.00000000000006.

4) * Оценка условной вероятности подписки при условии просмотра телевизора не реже раза в неделю оказалась равна 0.0000000000002.

Часть 7. Сравнение моделей.

7.1. Предсказательная сила.

Предсказательные силы линейно-вероятностной и пробит моделей как доли верных предсказаний были вычислены ранее, в пункте 3.6. Сделав соответствующие действия для логит-модели, пробит-модели с учетом гетероскедастичности и первого уравнения из системы бинарных пробит уравнений получаем результаты для сравнения в Таблице 9:

Linear probability	0.6672
Logit	0.6654
Probit	0.6664
Heteroskedasticity probit	0.6648
Binary probit model	0.6682

Таблица 9. Сравнение моделей по предсказательной силе.

Среди построенных нами моделей наибольшей предсказательной силой (Таблица 9) обладают линейно-вероятностная модель и модель из системы бинарных пробит уравнений, так как у них наибольшая доля верных предсказаний. Далее идет пробит-модель, у которой доля правильно предсказанных значений немного ниже. Оказалось, что среди наших моделей наименьшей предсказательной силой обладают пробит модель с учетом гетероскедастичности, так как она имеет наименьшую среди рассмотренных долю верных предсказаний.

7.3. Информационные критерии.

Еще одни способом выбрать лучшую из оцененных моделей является использование информационных критериев AIC и BIC. Информация о данных критериях для наших моделей подставлена в Таблице 10:

	AIC	*BIC*
Linear probability	6489.385	6535.006
Logit	6182.81	6221.913
Probit	6182.855	6221.958
Heteroskedasticity probit	6178.047	6230.185
Binary probit model	12071.57	12156.3

Таблица 10. Сравнение моделей по информационным критериям.

При сравнении моделей по информационным критериям (Таблица 10) необходимо выбирать модель с наименьшим в абсолютном выражении значением критерия. По критерияю Акаике лидирует пробит-модель с учетом гетероскедастичности с наименьшим значением для данного показателя среди оцененных моделей. Далее идет пробит модель, чье значения информационного критерия немного выше, чем для пробит модели с учетом гетероскедастичности. Что касаемо Байесовского информационного критерия, то здесь логит и пробит модели поменялись местами: у логит модели наименьшее значение данного критерия, а у пробит модели с учетом гетероскедастичности немного выше.