Пермский филиал федерального государственного автономного образовательного учреждения высшего образования   
«Национальный исследовательский университет   
«Высшая школа экономики»

*Факультет экономики, менеджмента и бизнес-информатики*

Можегова Анна Сергеевна

**Модульное задание №2 по дисциплине**

**«Классификация статистических данных»**

студента образовательной программы «Программная инженерия»

по направлению подготовки *09.03.04 Программная инженерия*

**Руководитель:**

Тихонова Арина  
 Михайловна

Пермь, 2023 год

Оглавление

[Введение 3](#_Toc134636511)

[Глава 1 Предварительное описание данных 5](#_Toc134636512)

[1.1 Экономическая интерпретация 5](#_Toc134636513)

[1.2 Характеристики по шкалам 5](#_Toc134636514)

[Глава 2 Дискриминантный анализ 6](#_Toc134636515)

[2.1 Метод принудительного включения 6](#_Toc134636517)

[2.2 Пошаговый дискриминантный анализ 13](#_Toc134636518)

[2.3 Метод принудительного включения после удаления выбросов 18](#_Toc134636519)

[2.4 Пошаговый дискриминантный анализ после удаления выбросов 22](#_Toc134636520)

[2.5 Целесообразность проведения дискриминантного анализа 27](#_Toc134636521)

[Глава 3 Построение дерева классификации 28](#_Toc134636522)

[3.1 Рассмотрение различных способов выбора зависимых и независимых переменных 28](#_Toc134636524)

[3.2 Интерпретация результатов 31](#_Toc134636525)

Введение

В представленной работе выполняется дискриминантный анализ и построение дерева решений методом CHAID. В качестве объекта исследования были взяты данные из прошлой модульной работы, которые были отфильтрованы, а именно столбец со среднемесячной заработной платой по полному кругу организация был исключен, поскольку в прошлой работе выяснилось, что он мало с чем классифицируется, а также были добавлены данные на вопросы «Является ли лидером российского туризма?» и «Регион входит в "Национальный экологический рейтинг"?». Таким образом для анализа собраны данные об оценке туристического потока, обороте общественного питания на душу населения, валовая добавленная стоимость туристической индустрии, численность размещенных лиц в коллективных средствах размещения, валовой региональный продукт на душу населения и число турфирм за 2022 год и показатели, которые отображают ответы на поставленные вопросы.

Дискриминантный анализ – это метод статистического анализа, который используется для выявления различий между группами объектов на основе нескольких признаков. Он позволяет определить наиболее значимые признаки, которые отличают объекты разных групп друг от друга. Дискриминантная модель, построенная с помощью этого метода, позволяет классифицировать наблюдения в соответствии с их характеристиками. Дискриминантный анализ работает с переменными и другими факторами, используемыми в качестве предикторов.

CHAID – один из методов анализа с помощью дерева решений. Дерево решений строится путем разбиения выборки на подмножества, каждое из которых соответствует определенному значению признака. Каждый узел дерева представляет собой тест на значение одного из признаков, а каждая ветвь – возможное значение этого признака. Листья дерева соответствуют решению, которое принимается на основе значений всех признаков, входящих в путь от корня до листа. Преимущества использования дерева решений в анализе данных заключаются в том, что оно позволяет легко интерпретировать результаты и принимать решения на основе наглядной структуры дерева.

Целью работы является проведение всех перечисленных анализов имеющихся данных.

Задачами, которые необходимо выполнить для достижения поставленной цели, являются:

* рассмотрение метода принудительного включения в дискриминантном анализе;
* рассмотрение пошагового дискриминантного анализа;
* удаление выбросов и интерпретация результатов для метода принудительного включения и пошагового дискриминантного анализа;
* описание целесообразности проведения дискриминантного анализа;
* построение дерева решений;
* описание результатов метода CHAID;
* описание целесообразности построения дерева решений;
* привести общий вывод о проделанной работе.

# **Глава 1 Предварительное описание данных**

Это глава, в которой производится описание показателей, рассматриваются различные характеристики по шкалам.

* 1. Экономическая интерпретация

Для своего исследования в качестве исходных данных я брала данные по регионам, а именно оценку туристического потока, оборот общественного питания на душу, валовая добавленная стоимость туристической индустрии, численность размещенных лиц в коллективных средствах размещения, валовой региональный продукт на душу населения и число турфирм

* 1. Характеристики по шкалам

Названия регионов является номинальной шкалой, ведь это данные, с которыми мы не может делать каких-либо арифметических операций, а вот все остальные, кроме «Является ли лидером российского туризма?» и «Регион входит в "Национальный экологический рейтинг"?»— это интервальные шкалы, так как есть возможность находить разницу между двумя величинами, таким образом, мы, например, можем рассмотреть разность турфирм 2х различных регионов. «Является ли лидером российского туризма?» и «Регион входит в "Национальный экологический рейтинг"?» являются бинарными данными.

# **Глава 2 Дискриминантный анализ**

В данной главе поэтапно показывается проведение дискриминантного анализа, для этого была использована среда SPSS, в которой есть встроенные функции.

1. 1. Метод принудительного включения

Рассмотрим дискриминантный анализ по методу принудительного включения, выберем группировку по столбцу «Является ли лидером российского туризма?»

Нужно учитывать, что несмотря на то, что функция классификации текущим методом встроенная нужно также указывать некоторые параметры в настройках, таким образом, была произведены изменения в «Статистика», «Классифицировать» и «Сохранить».

В результате обработки запроса получаем следующие сведения, например, анализ сводки обработки наблюдений, рассмотрим его на рисунке 2.1.1.

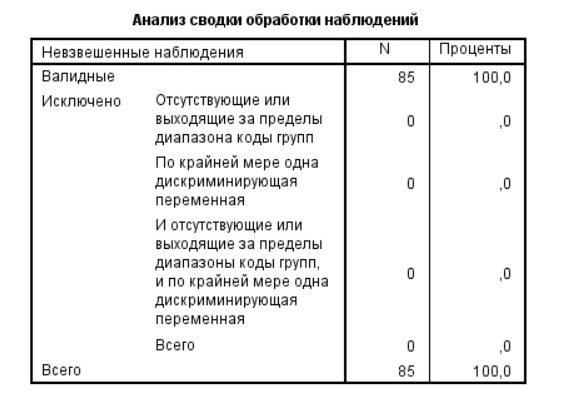
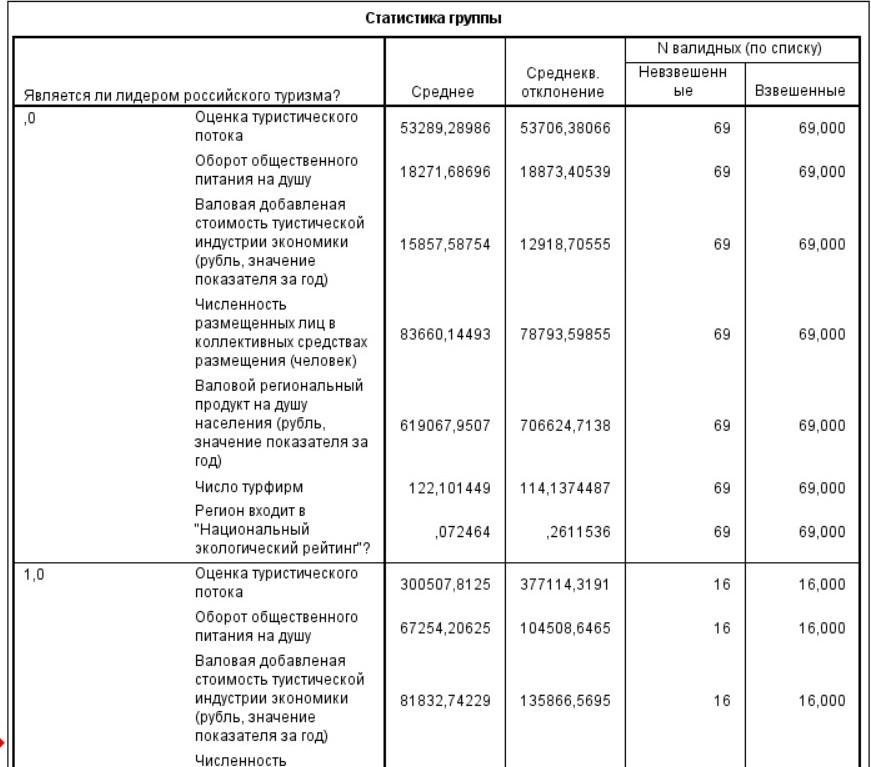
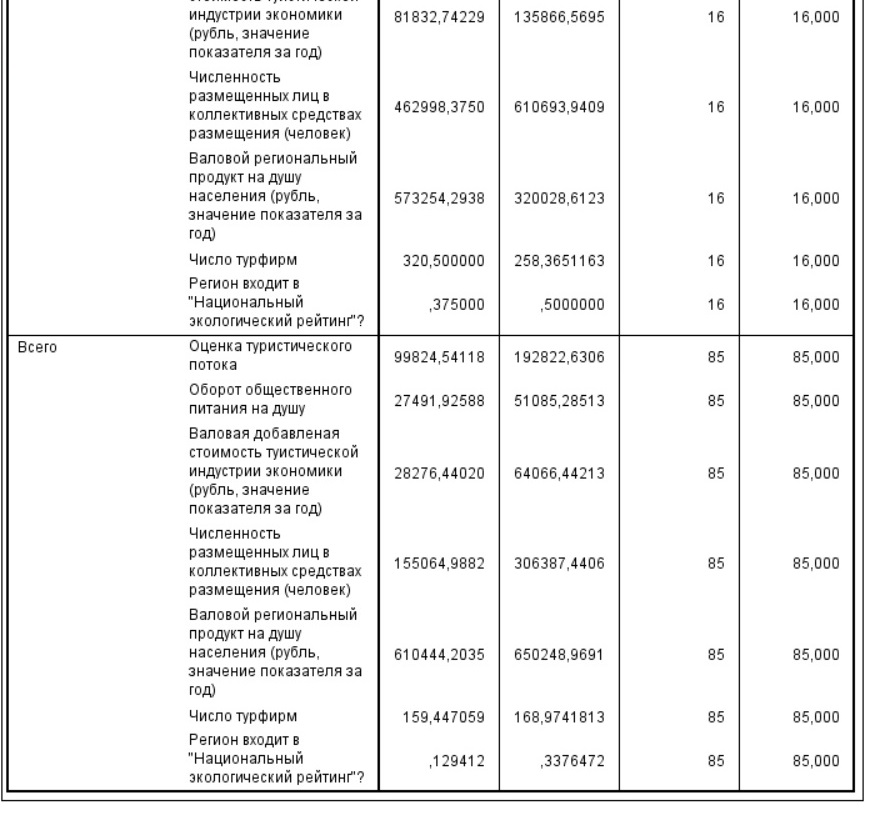


Рисунок 2.1.1 – Изображение таблицы анализа сводки обработки данных

Всего 85 наблюдения. Из сводки видно, что отсутствуют исключенные из анализа наблюдения, а это значит, что для всех наблюдений проведен дискриминантный анализ.

Далее идет статистика группы, здесь содержатся данные о средних значениях дискриминационных переменных в каждой исследуемой группе. Эти показатели дают общее представление о том, являются ли дискриминационные переменные отличительными признаками этих групп. Таблица приведена на рисунках 2.1.2 и 2.1.3

 Рисунок 2.1.2 – Первая часть сводной таблицы об статистике группы

 Рисунок 2.1.3 – Вторая часть сводной таблицы об статистике группы

По таблице видно, что средние показатели значительно отличаются, но стоит отметить, что неравенство средних не доказывает, что эти переменные являются отличительными признаками исследуемых групп. Необходимо установить статистическую значимость различий средних, проведя тест на равенство средних.

Для этого используем лямбду Уилкса, данное понятие представляет из себя критерий, который используют при проведении теста в результате он показывает значимо ли различаются между собой средние значения дискриминантной функции в исследуемых группах. Если лямбда меньше 0.05, значит средние двух групп значимо различаются, т.е. доказано наличие дискриминирующих особенностей этих переменных. В ином случае необходимо убирать предикторы из анализа. На рисунке 2.1.4 представлена таблица критерия равенства групповых средних, в которую включена лямбда Уилкса.

 Рисунок 2.1.4 –Критерии равенства групповых средних

Таким образом мы видим, что значимость у предиктора большая «Валовой региональный продукт на душу населения» и его необходимо удалить.

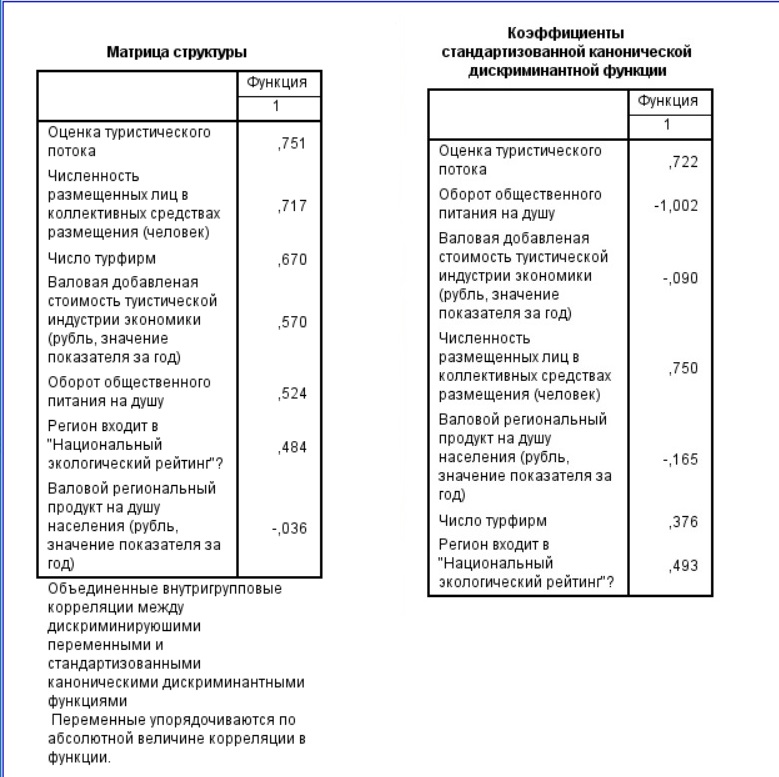
Далее идут таблицы «Собственные значения» и «Лямбда Уилкса» (см. рис. 2.1.5). В первой из них присутствуют понятия собственное значение и каноническая корреляция, собственное значение — это отношение межгрупповой суммы квадратов к внутригрупповой сумме квадратов, каноническая корреляция — это коэффициент между рассчитанными значениями дискриминантной функции и показателем принадлежности к группе; показывает силу взаимосвязи между предикторными переменными и группами. Во второй таблице лямбда Уилкса показывает, значимо ли различаются в двух группах средние значения дискриминантной функции. Если значимость меньше 0,05 – значимое различие.



Рисунок 2.1.5 –Собственные значения и лямбда Уилкса

Из выше сказанного делаем вывод, что силу взаимосвязи между предикторными переменными и группами достаточно сильная, как и отношение межгрупповой суммы квадратов к внутригрупповой сумме квадратов. По значению лямбды Уилкса видно, что между группами значимое различие.

Рассмотрим стандартизованные коэффициенты дискриминантных функций и матрицу структуры (см. рис 2.1.6).

 Рисунок 2.1.6 –Матрица структуры и коэффициенты стандартизованной канонической дискриминантной функции

Нужно отметить, что коэффициенты стандартизированной канонической дискриминантной функции используют как множители для нормированных переменных, т.е. переменных с нулевым средним и дисперсией, равной 1, так удается оценить относительный вклад каждой дискриминантной переменной в различие двух исследуемых групп. В матрице же показана сила связи дискриминантных переменных со стандартизованными значениями дискриминантной функции. Из показателей видно, что есть переменные, имеющие отрицательный вклад в различие двух групп, но все же положительных больше, кроме того, 2 из 4 принимают большое значение. Сила связи же по матрице в основном у всех выше 0.5, но также есть и одно отрицательное значение.

Следующим этапом является нестандартизированные (канонические) коэффициенты дискриминантной функции, представляющие из себя множители при заданных значениях переменных, входящих в дискриминантную функцию. Обратим внимание на рисунок 2.1.7.

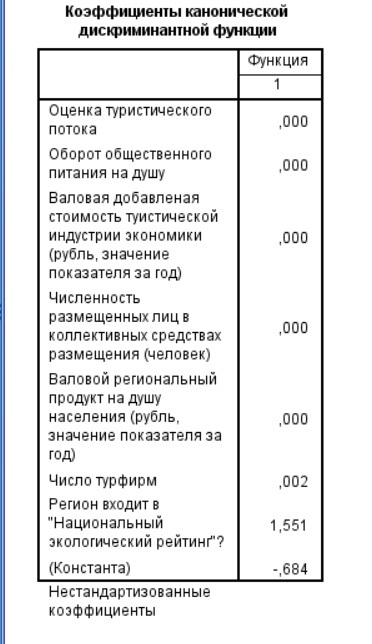


Рисунок 2.1.7 –Не стандартизированные коэффициенты канонической дискриминантной функции

Таким образом уравнение имеет вид , где все переменные по порядку отвечают за значения в таблице. На основе построенной дискриминантной модели можно сделать прогноз о том, какая часть областей/городов считает, что ее их присутствие в лидерах туризма важно при остальных показателях. Полученное значение D необходимо найти на диаграммах распределения значений дискриминантной функции.

Определим качество модели для этого есть таблица под названием «Функции в центроидах групп» (см. рис. 2.1.8). Таким образом четкость разделения исследуемых групп характеризуется расстоянием между центроидами – средними значениями дискриминантной функции в исследуемых группах. И соотношение таково, что чем больше расстояние, тем лучше разделены группы. Смотря на таблицу, то можно сказать, что разделены группы хорошо.

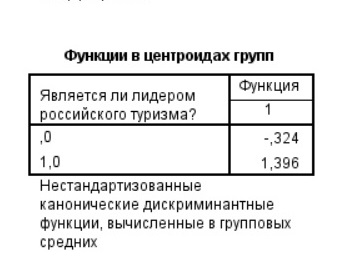
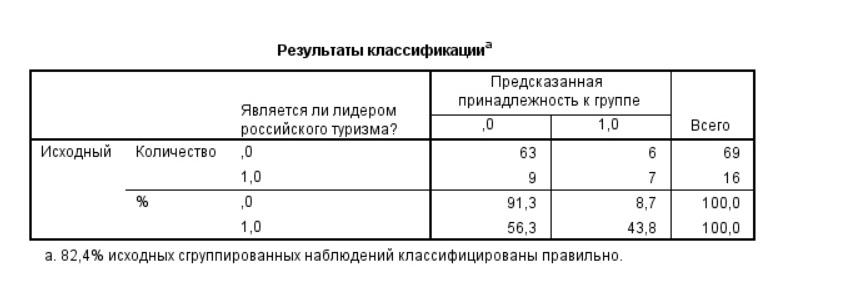


Рисунок 2.1.8 –Функции в центроидах групп

Кроме того, качество модели можно определить по результатам классификации, представленным на рисунке 2.1.9, где показано, что точность прогнозов составляет 82.4%

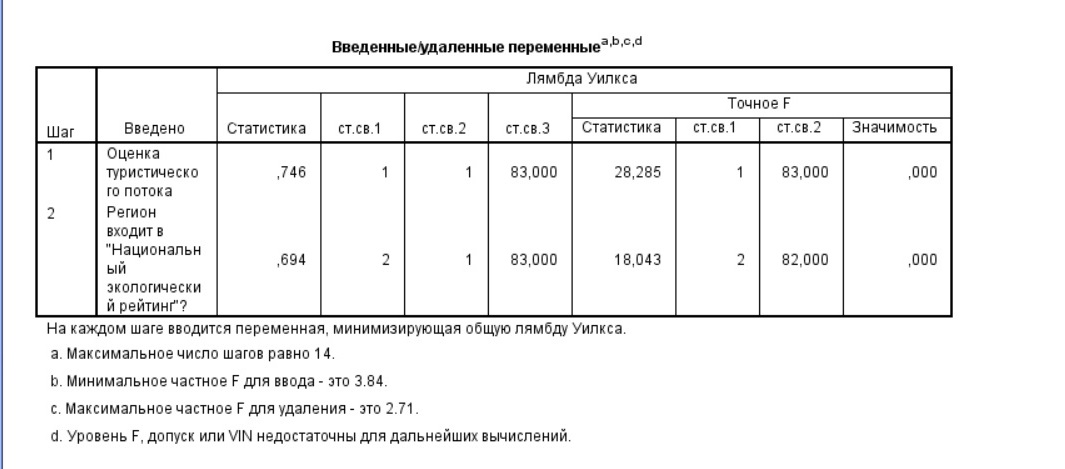
 Рисунок 2.1.9 –Результат классификации

* 1. Пошаговый дискриминантный анализ

Данный метод не будет описан подробно, поскольку цель подпункта — это сравнение результатов с методом принудительного включения.

Количество наблюдений также осталось 85, таблицы «Статистика группы», «Критерий равенства групповых средних», «Объединенные внутригрупповые матрицы» остались теми же.

Таблицы «Введенные/удаленные переменные», «Переменные для анализа» и «Переменные не для анализа» иллюстрируют пошаговый процесс составления дискриминантного уравнения. (см.рис.2.2.1)

 Рисунок 2.2.1 –Введенные/удаленные переменные

Появились новые таблицы, такие как «Переменные для анализа» и «Переменные не для анализа», не трудно понять по названию, что одна из них иллюстрирует те переменные, которые использоваться, а другая демонстрирует исключенные переменные. (см.рис.2.2.2, рис.2.2.3, рис.2.2.4)



Рисунок 2.2.2 –Переменные для анализа

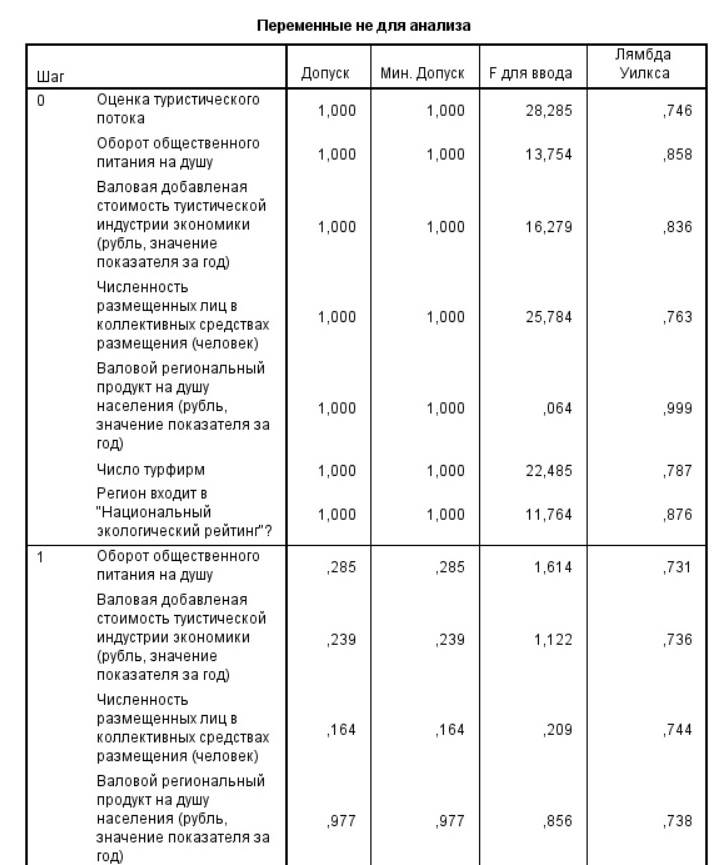


Рисунок 2.2.3 –Переменные не для анализа(часть первая)

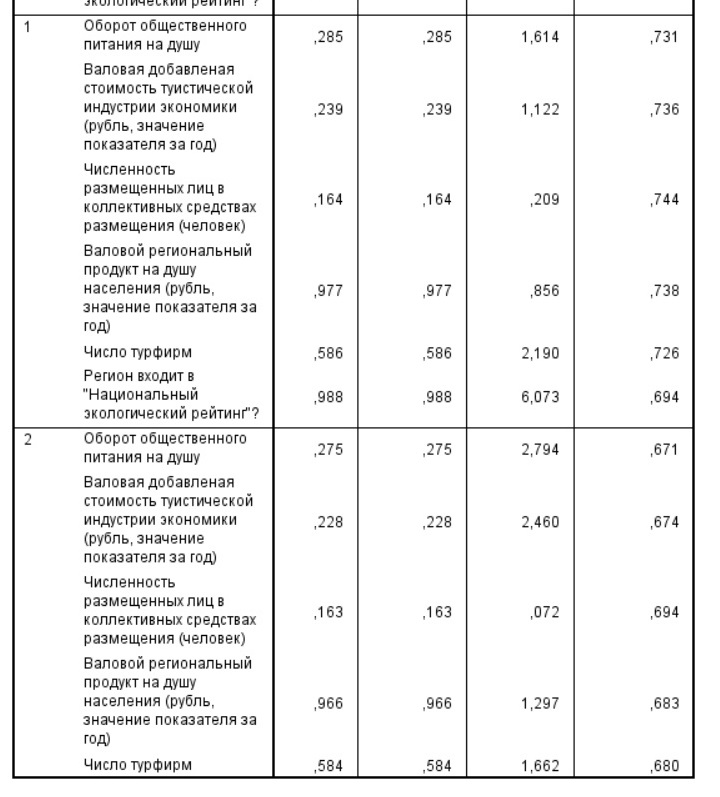


Рисунок 2.2.4 – Переменные не для анализа(часть вторая)

В таблице «Собственные значения» можно заметить, что каноническая корреляция изменилось с 0,614 до 0,553, т.е. уменьшилась, а это означает и уменьшение силы взаимосвязи между предикторными переменными и группами. По значению лямбды Уилкса все также между группами значимое различие. (см.рис.2.2.5)



Рисунок 2.2.5 –Собственные значения и лямбда Уилкса

Увеличился вклад оценки туристического потока. Таким образом уравнение имеет новый вид, а именно , где все переменные по порядку отвечают за значения в таблице «Коэффициенты канонической дискриминантной функции». Расстояние не изменилось (см. рис. 2.2.6)

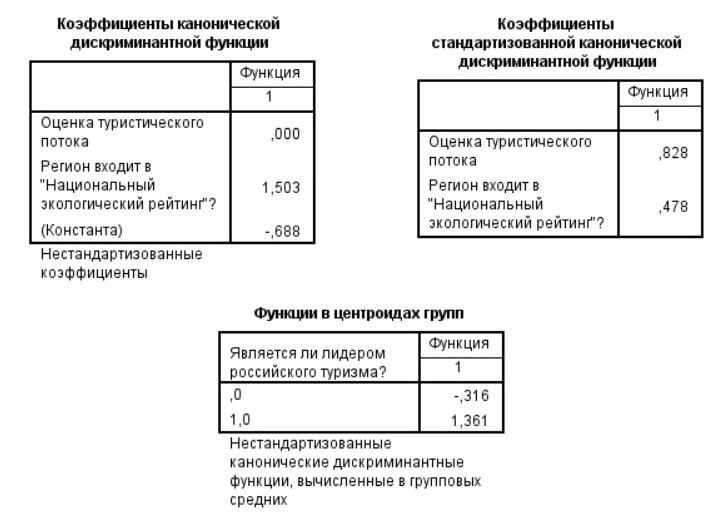


Рисунок 2.2.6 –Основные таблицы

Дискриминантные переменные сильнее связаны со стандартизованными значениями дискриминантной функции и это видно по матрице структуры. (см.рис.2.2.7)

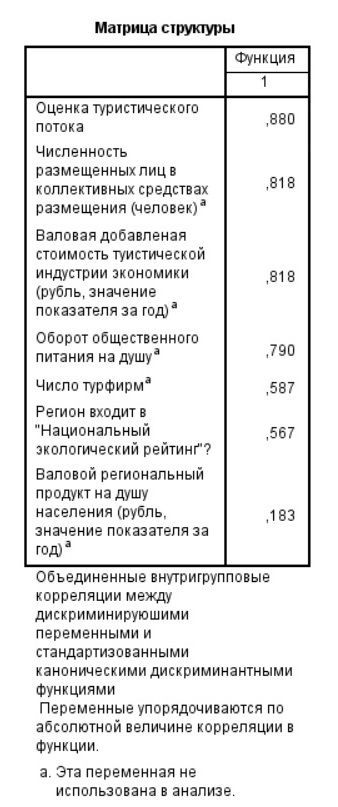


Рисунок 2.2.7 –Матрица структуры

Рассмотрим результаты анализа на рисунке 2.2.8, где можно увидеть точность модели. В данном случае она составляет 84,7%. Получается, что эта модель лучше предыдущей всего на 0.3% и они почти ничем не отличаются.

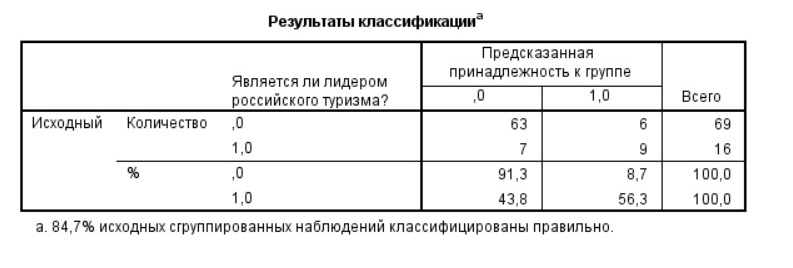


Рисунок 2.1.10 –Результат классификации

* 1. Метод принудительного включения после удаления выбросов

Рассмотрим результаты анализа после удаления выбросов, возможно, это улучшит модель.

По таблице «Анализ сводки обработки наблюдений» на рисунке 2.3.1 также видно, что все наблюдения попали в анализ.

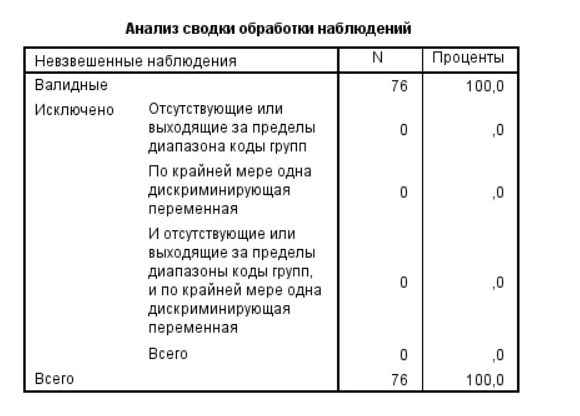
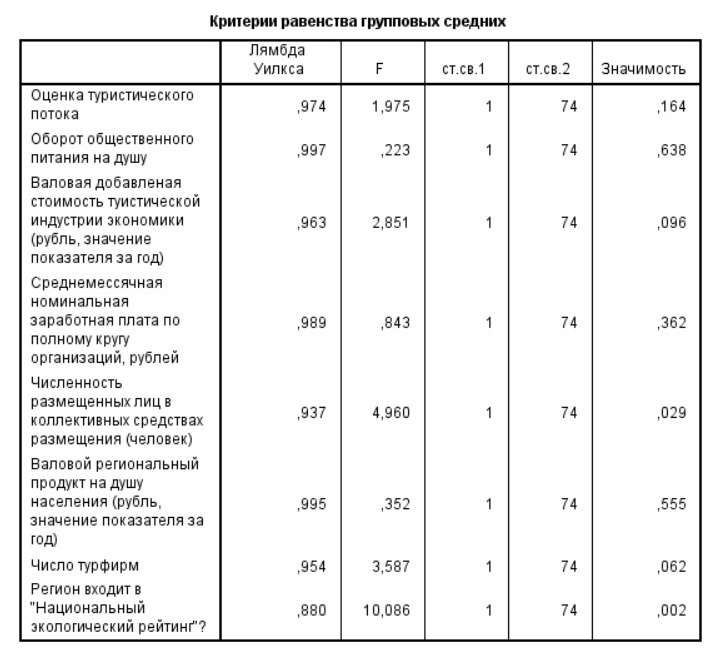


Рисунок 2.3.1 –Анализ сводки

Далее идет статистика группы, по этой таблице было видно, что средние показатели значительно отличаются, но мы рассмотрим значимость лямбды Уилкса на рисунке 2.3.2

 Рисунок 2.3.2 –Критерии равенства групповых средних

Ранее мы получали такой результат, что значимость была большой лишь у предиктора «Валовой региональный продукт на душу населения», сейчас значимость увеличилась почти у всех параметров и их исключением являются «Численность размещенных лиц в коллективных средствах размещения» и «Регион входит в Национальный экологический рейтинг».

Посмотрим на собственные значения и лямбду Уилкса, по полученным значениям видно, что сила взаимосвязи между предикторными переменными и группами осталась достаточно сильной, как и отношение межгрупповой суммы квадратов к внутригрупповой сумме квадратов. А вот значение лямбды Уилкса есть изменения поскольку коэффициент указывает на то, что между группами не значимое различие. (см.рис 2.3.3)

 Рисунок 2.3.3 –Собственные значения и лямбда Уилкса

Посмотрим на силу связи дискриминантных переменных со стандартизованными значениями дискриминантной функции в таблице «Матрица структуры», ранее в ней было лишь одно отрицательное значение, сейчас же их два, кроме того, в целом все значения стали меньше. По таблице «Коэффициенты стандартизованной канонической дискриминантной функции» делаются выводе о вкладе каждой дискриминантной переменной в различие двух исследуемых групп, отсюда видно, что до удаления другие переменные имели отрицательный вклад, например, до этого у оценки туристического потока было значение 0.722, а сейчас стало –0.298. Вклад переменных значительно изменился. (см.рис.2.3.4)



Рисунок 2.3.4 –Матрица структуры и коэффициенты стандартизованной канонической дискриминантной функции

Переходим к самому интересному, а именно к виду уравнения канонической дискриминантной функции, до этого были представлены два вида: и . Теперь уравнение имеет следующий вид .(см. рис 2.3.5)

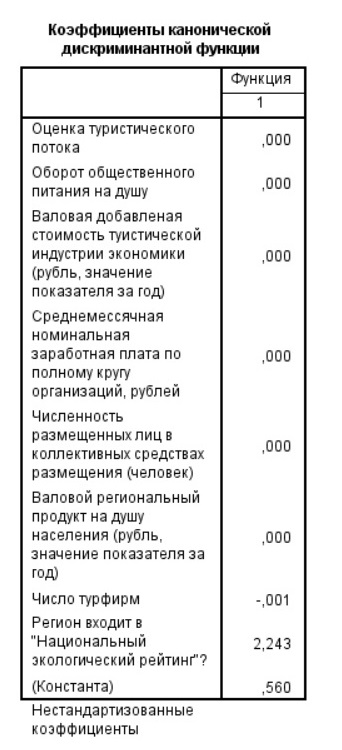


Рисунок 2.3.5 –Коэффициенты канонической дискриминантной функции

Функции в центроидах групп слегка изменились, оба значения увеличились, первый стало больше на 0.079, а второй на 0.038. В данном случае точность модели составляет 82,9%. Получается, что эта модель чуть хуже, тех, что получились с данными, в которых есть выбросы. (см. рис. 2.3.6)



Рисунок 2.3.6 –Оценка качества

* 1. Пошаговый дискриминантный анализ после удаления выбросов

Смотря на результаты метода «Принудительного включения» после удаления выбросов модель имеет меньшую точность, теперь рассмотрим какие результаты получаться пошаговым методом.

Как и во все прошлые разы из сводки видно, что отсутствуют исключенные из анализа наблюдения, а это значит, что для всех наблюдений проведен дискриминантный анализ. (см. рис 2.4.1)

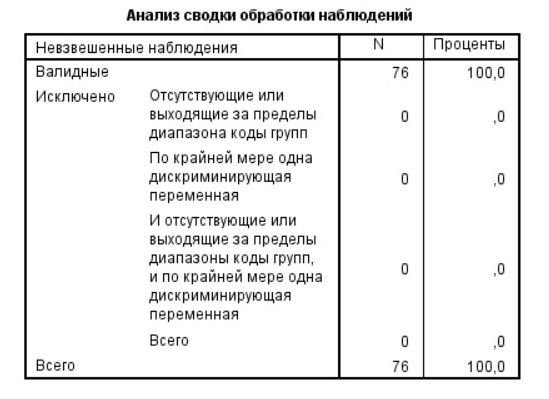


Рисунок 2.4.1 –Анализ сводки обработки наблюдений

Как уже известно по таблице «Критерии равенства средних» по значимости определяются предикторы, которые необходимо исключать из анализа, в предыдущем примере исключением были «Численность размещенных лиц в коллективных средствах размещения» и «Регион входит в Национальный экологический рейтинг», сейчас же исключением является «Регион входит в Национальный экологический рейтинг». Таблица продемонстрирована на рисунке 2.4.2.

 Рисунок 2.4.2 –Критерии равенства групповых средних

Рассмотрим, как изменились таблицы «Переменные для анализа» и «Переменные не для анализа». (см.рис.2.4.3, рис.2.4.4 и рис.2.4.5)

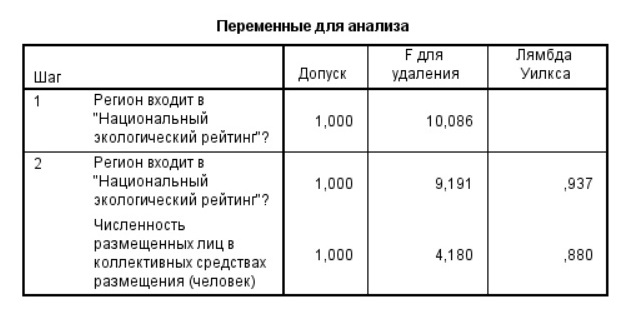
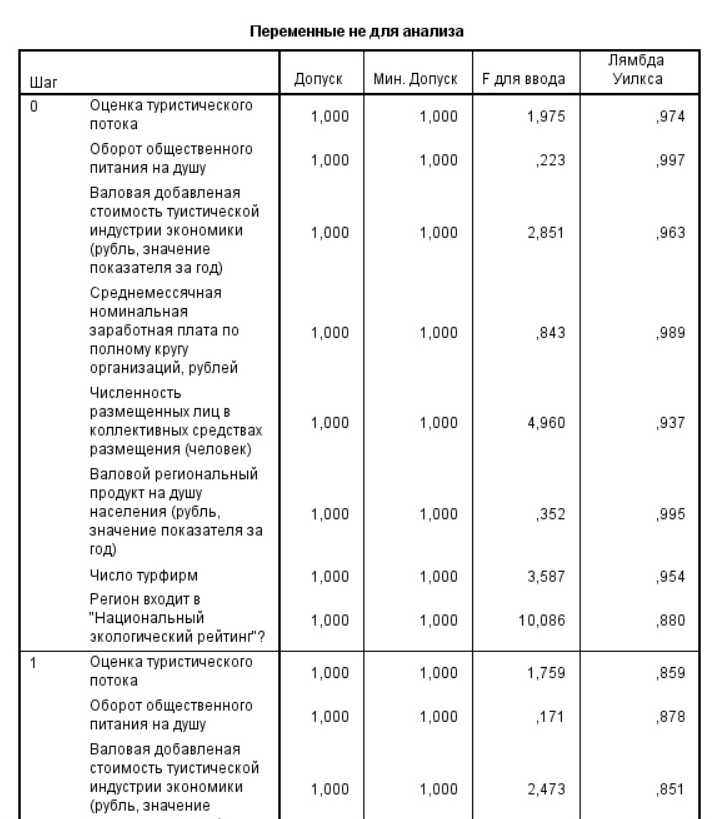


Рисунок 2.4.3 –Критерии равенства групповых средних

Ранее переменными для анализа было 2 переменные: «Оценка туристического потока» и «Регион входит в Национальный экологический рейтинг», после удаления выбросов получилось так, что переменными для анализа стали «Регион входит в Национальный экологический рейтинг» и «Численность размещенных лиц в коллективных средствах размещенния»

 Рисунок 2.4.4 –Критерии равенства групповых средних

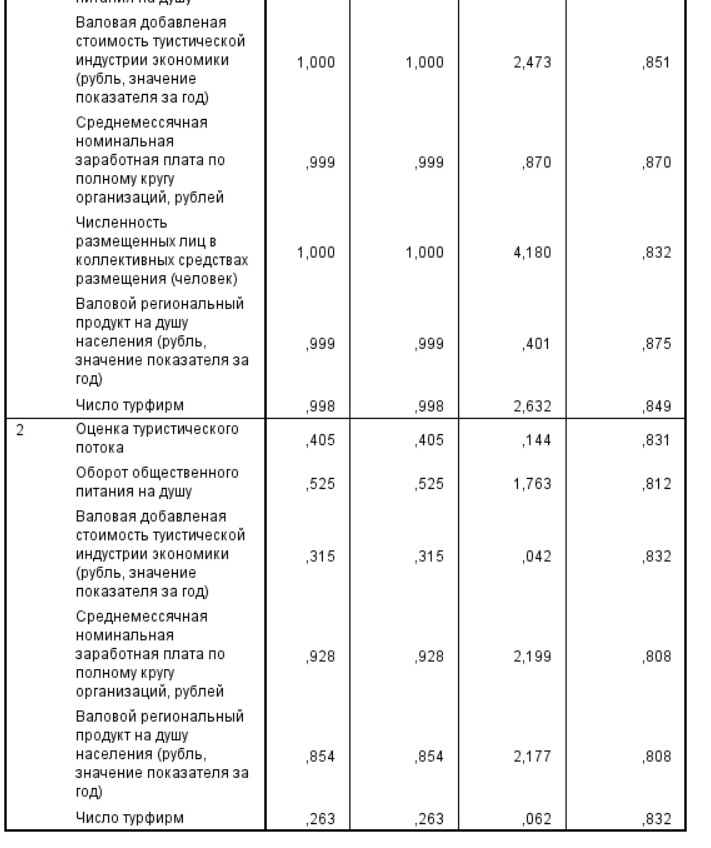


Рисунок 2.4.5 –Критерии равенства групповых средних

Что относится к таблице «Переменные не для анализа», то можно увидеть, что поменялись значения у столбцом «F для ввода» и «лямбда Уилкса», но это связано лишь с тем, статистический критерий F используется для определения значимости различий между группами переменных. Этот критерий показывает, насколько велика разница между группами переменных и является одним из факторов, которые могут быть использованы для выбора наиболее значимых переменных для включения в модель дискриминантного анализа. В общем, чем выше значение F, тем более значимыми считаются различия между группами переменных, кроме того, данный критерий изменился поскольку изменились и сами данные (были удалены выбросы), и различия изменились.

В таблице «Собственные значения» можно заметить, что каноническая корреляция изменилось с 0.553 до 0.409, т.е. уменьшилась, а это означает и уменьшение силы взаимосвязи между предикторными переменными и группами. По значению лямбды Уилкса все также между группами значимое различие. (см.рис.2.4.6)

 Рисунок 2.4.6 –Собственные значения и лямбда Уилкса

Посмотрим на таблицу «Матрица структуры», в ней также осталось два отрицательных значения и в целом все показатели снизились еще больше, если сравнивать их с результатами этого же метода до удаления и с результатами метода принудительного включения до удаления. Вклад «Регион входит в Национальный экологический рейтинг» увеличился на 0.339, а вклад от «Численности размещённых лиц в коллективных средствах размещения» составляет 0.568, отрицательных вкладов в различие двух исследуемых групп нет. (см.рис.2.4.7)

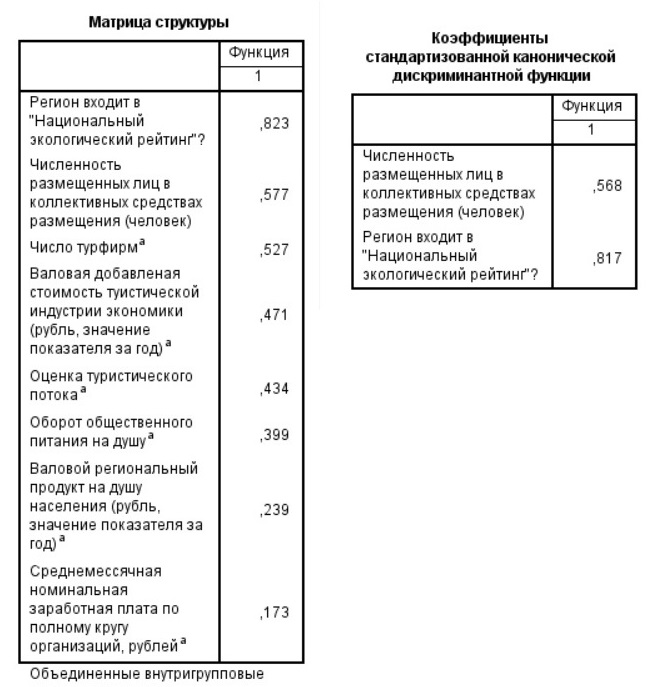


Рисунок 2.4.7 –Матрица структуры и коэффициенты стандартизованной канонической дискриминантной функции после удаления выбросов

Уравнение принимает следующий вид , где x-численность размещенных лиц в коллективных средствах размещения, а y-это входит ли регион в Национальный экологический рейтинг. Значения функции в центроидах также поменялась, первый показатель увеличился на 0.134, а второй уменьшился на 0.285(см.рис.2.4.8)

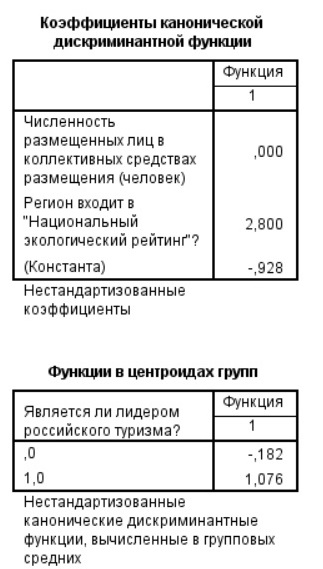


Рисунок 2.4.8 –Коэффициенты для составления уравнения и расстояние между группами

В результате анализа точность модели составляет 80,3%, что является самым маленьким значением из всех полученных. (см.рис.2.4.9)

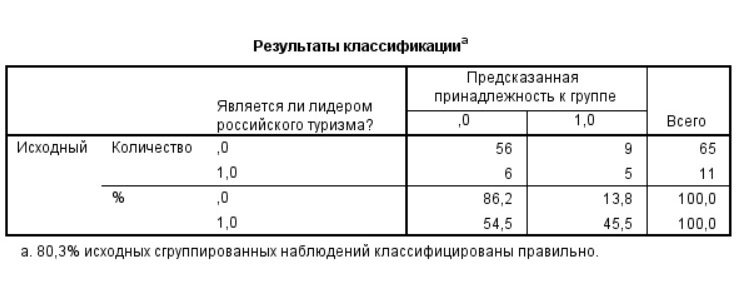


Рисунок 2.4.9 –Результат классификации

* 1. Целесообразность проведения дискриминантного анализа

Целесообразность проведения дискриминантного анализа можно оценить, используя следующие критерии:

* наличие различий между группами. Если известно, что группы имеют различия в характеристиках, то проведение дискриминантного анализа может быть целесообразным;
* наличие достаточного количества наблюдений для каждой группы. Для проведения дискриминантного анализа необходимо иметь достаточное количество наблюдений для каждой группы, чтобы результаты были надежными;
* наличие нормального распределения данных. Дискриминантный анализ предполагает нормальное распределение данных, поэтому если данные не соответствуют этому требованию, то результаты анализа могут быть неточными;
* отсутствие сильной корреляции между переменными. Если переменные сильно коррелированы, то они могут вносить дублирующую информацию в модель, что может привести к неверным выводам.

Таким образом, смотря на изначальные данные можно сказать, что проведения анализа было сомнительным поскольку между показателями существует большая корреляция, а также большинство из них не имеет нормального распределения. Даже после удаления выбросов ситуация не сильно изменилась. Тем не менее качество полученный моделей достаточно высоко.

# **Глава 3** **Построение дерева классификации**

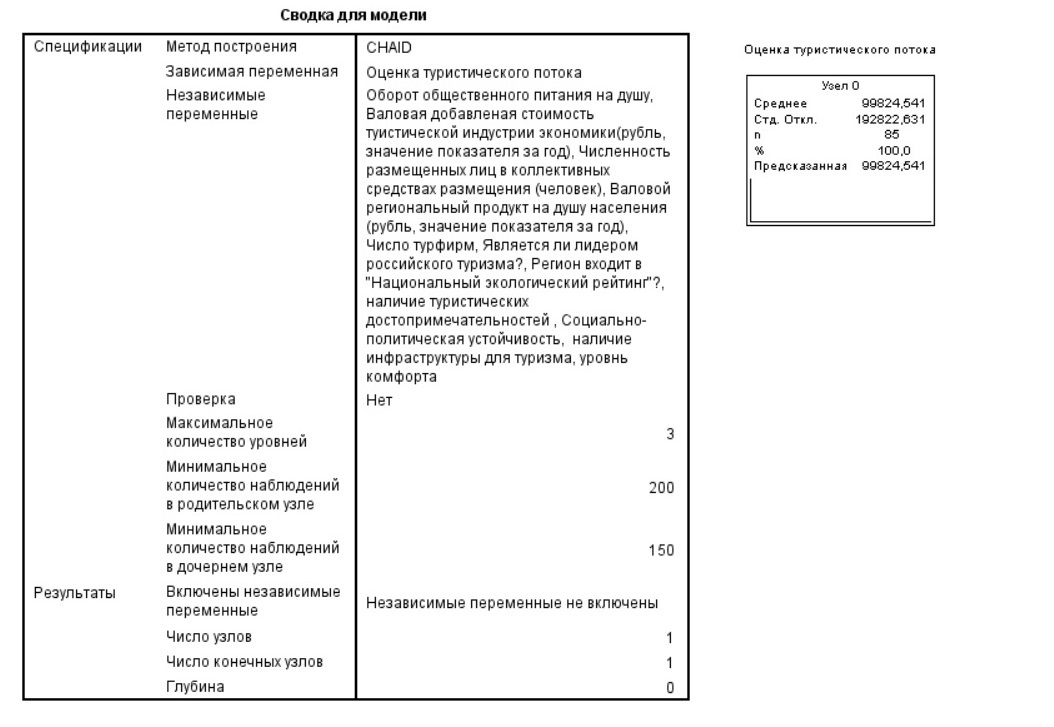
Глава 3 посвящена попыткам построения деревьев решений по методу CHAID. Этот метод имеет под собой практичный и быстрый поиск взаимосвязи между предикторными переменными и категориальным откликом. Его лучше выбирать в том случае, когда имеется много потенциальных предикторов, которые сложно анализировать посредством таблиц сопряженности, из-за их большого количества. Результатом приводится в виде древовидной структуры, которую получают с помощью иерархической сегментации данных.

Основная задача исследования с помощью деревьев решений заключается в том, чтобы используя заданный набор наблюдений (называемый обучающей выборкой), уловить скрытые статистические закономерности в данных (как одни случайные (независимые) характеристики влияют на интересующую нас (зависимую) характеристику) и построить модель зависимости в виде дерева решений.

1. 1. Рассмотрение различных способов выбора зависимых и независимых переменных

Предоставленное техническое задание предусматривает, что в качестве независимой переменной будет дискретная шкала, в моих данных такими показателями являются: оценка туристического потока, численность размещенных лиц в коллективных средствах размещения (человек), число турфирм и наличие туристических достопримечательностей. Рассмотрим варианты, когда каждый из этих столбцов выступит в качестве зависимой переменной.

В результате, когда зависимой переменой является оценка туристического потока, то мы получаем сводные таблицы, приведенные на рисунках 3.1.1,3.1.2.

 Рисунок 3.1.1 –Результат дерева решений

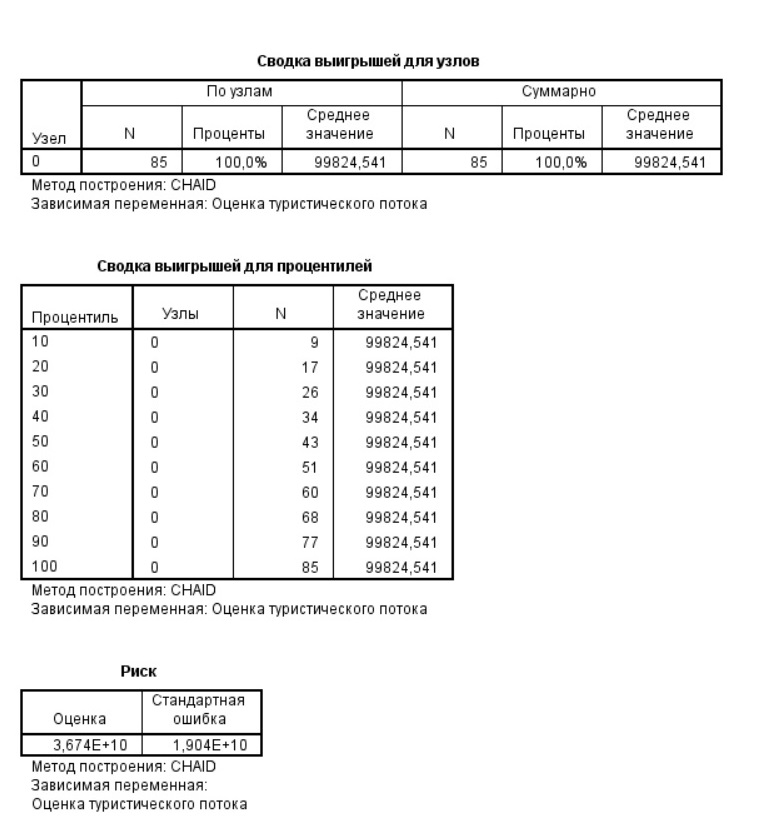


Рисунок 3.1.2 –Результат дерева решений

По приведенным изображениям выше видно, что дерево решений имеет лишь один узел, содержащий зависимую переменную. Данный метод работает по следующему принципу:

1. Поиск самого сильного фактора, который наибольшим образом объясняет различия.
2. Перебор всех заданных предикторов, поиск комбинаций значений и нахождение лучшего результата (который максимизирует различия). Выделение групп по найденному лучшему результату.
3. Повторение процесса (пунктов 1 и 2) с целью нахождения оптимального решения для второго уровня и т.д. для всех возможных уровней.
4. Представление результатов в виде дерева решений.

Получается, что на пункте 2 дерево решений прекращало строить свои узлы, поскольку не находилось подходящего результата.

Ниже приведены объеденные результаты остальных попыток построения дерева решений.

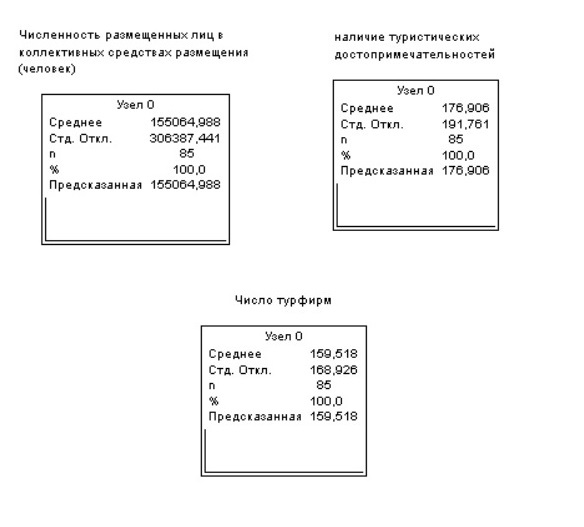


Рисунок 3.1.3 –Результаты дерева решений

Отсюда следует, что несмотря на выбор зависимой переменной результат был один, также ради эксперимента в виде зависимой переменной выбирались параметры, у которых значения принимают ответ на поставленный вопрос «Да/Нет» или «Среднее/Высокое», но это никак не повлияло на вид дерева. А если судить по таблице «Риск» о значении оценки, то можно сказать смело, что ошибка классификации велика и непонятно в каких процентных диапазонах она находится.

* 1. Интерпретация результатов

Дерево решений может иметь только один узел, если все наблюдения в выборке имеют одинаковое значение зависимой переменной. Например, если мы строим дерево решений для предсказания того, кто из клиентов купит продукт, а все клиенты в нашей выборке уже сделали покупку, то все наблюдения будут иметь значение "да" в качестве зависимой переменной. В этом случае дерево решений не будет иметь никакой информативности и не будет использоваться для прогнозирования.

Чтобы решить данную проблему, можно попробовать изменить выборку или добавить новые переменные, чтобы создать разнообразие в значениях зависимой переменной. Если это невозможно, то можно использовать другие методы анализа данных, например, логистическую регрессию или метод опорных векторов. Если дерево классификации SPSS имеет только один узел, то это может быть связано с недостаточной сложностью модели или недостаточным количеством данных для построения более сложной модели. Рекомендуется проверить наличие достаточного количества данных и использовать методы оценки качества модели для выбора оптимальной сложности дерева. Также может быть полезным проверить правильность выбора предикторов и их значимость для прогнозирования зависимой переменной.

Для проверки правильности выбора предикторов и их значимости для прогнозирования зависимой переменной можно использовать следующие методы:

* анализ корреляций между предикторами и зависимой переменной. Если предикторы сильно коррелируют друг с другом, то это может привести к переобучению модели и ухудшению ее качества. Также следует обратить внимание на корреляцию каждого предиктора с зависимой переменной, так как это показывает их значимость для прогнозирования;
  + - использование методов отбора признаков, таких как метод главных компонент или регрессионный анализ. эти методы позволяют определить наиболее значимые предикторы и убрать из модели ненужные;
    - построение моделей с разным набором предикторов и сравнение их качества. это позволяет определить, какие предикторы вносят наибольший вклад в прогнозирование зависимой переменной;
    - использование статистических тестов, таких как t-тест или анализ дисперсии, для проверки значимости каждого предиктора отдельно;
    - визуализация данных с помощью диаграмм рассеяния или графиков зависимости для выявления закономерностей и выбросов, которые могут влиять на качество модели;

В своем анализе я добавляла новые данные, т.е. независимые переменные, которые коррелировались значительно меньше представленных, но, к сожалению, результат дерева решений никак не поменялся. Построение моделей с разными предикторами также использовалось.

# **Вывод**

По итогам исследовательской работы были достигнуты все цели и выполнены все задачи. Были рассмотрены два статистических метода, их результатами не самые благополучные несмотря на то, что было представлено несколько вариантов построения дерева решений, а также дискриминантный анализ рассматривался как по методу принудительного включения, так и по пошаговому методу, также предоставлены результаты, когда в данных имелись выбросы и были удалены, но значительного влияния на анализ это не сделало. Для построения дерева решений были представлены вариации зависимых и независимых переменных, а также добавлялись дополнительные менее коррелируемые показатели, однако это также не улучшило представление дерева. Такие проблемы возникли по ряду причин, но основной является большая корреляция между переменными.

Были получены 4 вида уравнений для дискриминантного анализа, а также несколько представлений деревьев решений, которые имели лишь корень из-за некоторых причин, а именно есть предположение, что корреляция и отсутствие нормального распределения данных помешало провести анализ. В дополнении можно сказать, что точность моделей в дискриминантном анализе была не ниже 80%, а это говорит о том, что качество моделей высоко. А если судить по таблице «Риск» о значении оценки дерева решений, то можно сказать, что ошибка классификации велика и непонятно в каких процентных диапазонах она находится.

Литература и используемые информационные источники

Данные для исследования собирались на сайте Росстат:

<https://rosstat.gov.ru/folder/10705>

Приложение А

**Все операции в SPSS.**