Sehr geehrte Damen und Herren, in meiner heutigen Präsentation geht es um die Analyse von Süßwaren auf deren Beliebtheit und auf Basis dieser Analyse eine Empfehlung für die Eigenschaften einer neuen Süßigkeit.

Mein Vortrag ist in drei Teile gegliedert:

Zunächst werde ich kurz auf den Überblick der Daten eingehen. Dieser Datensatz enthält 85 unterschiedliche Süßwaren und 12 Eigenschaften. Das Kreisdiagramm zeigt die Anzahl der Süßwaren mit verschiedenen Eigenschaften. Auffällig entfällt der Hauptteil auf die Süßigkeiten, die in einer Tüte kommen. Der Anteil beträgt ungefähr 22%, gefolgt von Frucht und Schokoladen. Das Histogramm auf der rechten Seite zeigt, dass sich die meisten Gewinnprozentsatz auf etwa 40% konzentrieren.

Im zweiten Teil meiner Präsentation werde ich Ihnen das Heatmap bzw. die wichtige Korrelation zwischen Eigenschaften vorstellen. Durch Heatmap kann ich erfahren, dass die Korrelationen zwischen Schokolade und Gewinnprozentsatz am größten sind, was bedeutet, dass die Menschen Süßigkeiten mit Schokolade mögen, gefolgt von Erdnuss.

Zweitens, die Korrelationen zwischen Schokolade und fruchtige Süßigkeiten sind am geringsten, was bedeutet, dass Süßigkeiten in den meisten Fällen nicht mit Schokolade und Früchten gemischt werden. Außerdem werden nicht so viele Schokolade in einer Tüte verkauft.

Drittens, Schokolade werden oft von Tafel begleitet.

Viertens, Tafel sind teuer, aber die Kunden bevorzugen sie immer noch.

Fünftens, fruchtige Süßigkeiten werden oft in Tüten verpackt oder sind hart und billig.

Sechstens, Menschen mögen keinen fruchtigen oder harten Süßigkeiten oder verpackt in eine Tüte.

Siebentens, Süßigkeiten mit höherem Preis haben tatsächlich einen höheren Gewinnprozentsatz. Dies könnte daran liegen, dass teurere Süßigkeiten die Erwartungen der Kunden erhöhen.

Achtens, Alle Eigenschaften im Datensatz zeigen eine Korrelation von mehr als 0,1 mit dem Gewinnprozentsatz, deshalb ist jede Eigenschaft für den Gewinnprozentsatz relevant.

Danach komme ich auf die Zusammenfassung der Analyse jeder Süßware zu sprechen.

- Schokolade, Karamell, Erdnuss, Waffel und Tafel sind die wichtigen Merkmalen für eine Süßware mit hohem Gewinnprozentsatz.

- Die meisten Süßigkeiten sind ohne Nougat, und es gibt nur kleinen Unterschied, ob Nougat in der Süßware enthalten ist.

- (Je süßer die Süßware ist, desto beliebter ist sie.)

- Wenn die Süßigkeiten mehr als 60% Gewinnprozentsatz haben, muss die Süßware Schokolade enthalten, dann wird eine von Erdnuss und Karamell oder beide gewählt.

- Süßigkeiten ohne Schokolade sollten fruchtig, weich, billig und in einer Tüte verpackt sein.

Nun spreche ich über Hauptkomponentenanalyse. Mithilfe dieser Methode wird die Dimensionen reduziert und die bedeutenden Eigenschaften entnommen. Schokolade und Tafel sind die wichtigsten. Hart, fruchtig und Verpackung in einer Tüte hat die entgegengesetzte Richtung für Gewinnprozentsatz, daher sollten diese Eigenschaften in den Süßigkeiten nicht enthalten werden. Aber die erklärte Varianz der ersten drei Hauptkomponenten beläuft sich nur auf etwa 65% und so viele Informationen sind verloren. Das heißt, zwischen den Merkmalen besteht keine Multikolinearität. Das Ergebnis ist nicht so gut, deshalb dient dieses Ergebnis nur als Referenz.

Im folgenden möchte ich Ihnen Random forest regressor und XGB regressor Modelle kurz darlegen, um die relative Bedeutung der einzelnen Merkmale für den Gewinnprozentsatz zu bekommen. Weil die unabhängige variable kontinuierlich ist, ist diese Datensatz nicht für Klassifikation Methoden geeignet. In Random forest regressor sind Schokolade, Erdnuss und Tafel die wichtigen Eigenschaften. Aber in XGB regressor ist nur Schokolade wesentlich. Der Effekt ist nicht gut und Überanpassung ist erschienen, weil die Daten zu wenig sind.

Am Ende möchte ich zusammenfassen.

1) Schokolade ist die wirksamste Komponente, die sich auf die Beliebtheit auswirkt. Im Gegensatz dazu geht die fruchtigen Süßigkeiten in einer anderen Richtung.

2) Die fruchtigen und harten Süßigkeiten sollten möglichst nicht hergestellt werden.

3) Süßigkeiten, die Schokolade und Karamell oder Erdnuss in Form von Tafel oder Waffel enthalten, werden wahrscheinlicher verkauft.

4) Süßigkeiten ohne Schokolade sollten fruchtig, weich, billig und in einer Tüte verpackt sein.

5) Wenn wir wollen, dass die Süßware mehr als 60% Gewinnprozentsatz hat, müssen Süßwaren Schokolade enthalten, dann wird eine von Erdnuss und Karamell oder beide gewählt.

6) Ich empfehle, weiche Tafel-Süßwaren mit Schokolade und ohne Früchte, separate Verpackung, dann wählen Sie eine oder mehrere aus Erdnuss, Waffel und Karamell und fügen sie hinzu. Der Preis könnte auf mehr als 60 Prozent erhöht werden. Ich vermute, dass die Süßwaren mit oben genannten Zutaten bereits sehr süß sind, so dass es nicht notwendig ist, Zuckerprozent zu kontrollieren.

Vielen Dank für Ihre Aufmerksamkeit.

Die Ridge-Regression ist eine Methode zur Modelloptimierung, die zur Analyse von Daten mit Multikollinearität verwendet wird. Diese Methode führt eine L2-Regularisierung durch. Wenn das Problem der Multikollinearität auftritt, sind die kleinsten Quadrate unverzerrt, und die Varianzen sind groß, was dazu führt, dass die vorhergesagten Werte weit von den tatsächlichen Werten entfernt sind.

Die Random Forest Regression ist ein überwachter Lernalgorithmus, der eine Ensemble-Lernmethode für die Regression verwendet. Es ist leistungsstark und genau.

XGBoost steht für "Extreme Gradient Boosting" und ist eine Implementierung des Gradient-Boosting-Baum-Algorithmus. XGBoost ist ein beliebtes Modell für überwachtes maschinelles Lernen mit Merkmalen wie Berechnungsgeschwindigkeit, Parallelisierung und Leistung.

Besonders wichtig/interessant/ … ist es, dass …

Das ist sehr wichtig/interessant/ …, da/weil …

Ich möchte besonders darauf hinweisen, dass …

Zum Schluss werde ich Sie hinweisen.