Revenue Leakage Analysis Report for Telus

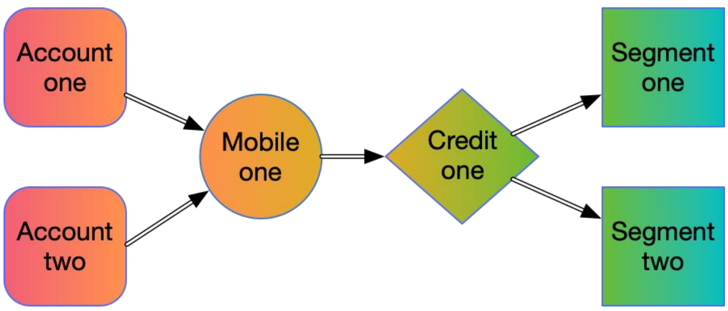
1. Introduction

Revenue Leakage Analysis (RLA) or Revenue Assurance (RA) represents a top priority function for most of the telecommunication operators worldwide [1]. Revenue leakage, if not prevented, depending on the severity of the leakage affecting their profitability and continuity, could cause a significant revenue loss of an operator. Detecting and preventing revenue leakage is a key process to assure telecom systems and processes efficiency, accuracy and effectiveness. But it is not an easy project to track back all the sources and root causes of the leakage issue considering millions or even billions of records existed in tiered product plans and flat rates [2]. Therefore, mining these big data effectively, predicting the leakage accurately and automating the analysis process would greatly increase performance, ease the auditing process, save operators revenues, provide better analytical experience, better management of data, more accurate reports and leads to an informed future decision making. The machine learning methods, such as the deep learning, random forest and xboost, have great potential to deal with this difficult problem. It should be noted that, in this report, these methods won't be used because the data prediction and classification are not required.

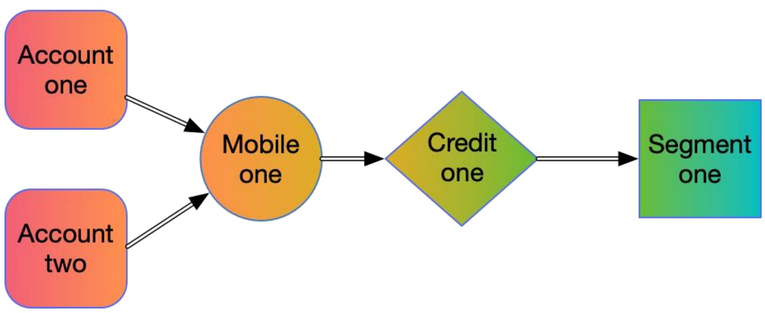
In this project, the objective is to find the relationship between false credit and revenue leakage. Then we can give some effective recommendations. The false credit means that the TELUS representatives don't adhere to the credit policy and give more credit than the allowed maximum credit to the customers. There are **four main questions** to answer, and I will answer these questions in four sections. The data source is an Excel document with four worksheets: Read me, Revenue_overage, Segment and Credits. The skills used in this project involve data extraction, data cleaning, data transformation, data analysis and data visualization. The language used in this project is python. The **ReadmeFirst file** in the folder shows the detailed information about the function of each python module. It also presents other information, such as package usage (pandas, openpyxl, numpy, etc.) and IDE. This report will show the discussion and results.

2. Initial Data Analysis

The data in the Excel document involve three relational tables: **Revenue_overage**, **Segment** and **Credits**. In order to answer the first question, these relational tables need to be joined. It can be easily found that the Revenue_overage table has the keys: account number and phone number; Segment table has the key: account number; Credits table has the key: phone number. Therefore, the Revenue_overage table is the bridge to join the three tables. Before we group the data in each table (groupby in pandas), the relationship between account number and phone number should be clarified first. Otherwise, duplicated calculation would be generated during the table joining process. It is found that their relationship is many-to-many (data_initial_analysis.py module). There are 25 phone numbers corresponding to two account numbers. There are 1158 account numbers corresponding to at least two phone numbers. As for the 25 phone numbers, it would lead to duplicated calculation for credit when we join Revenue_overage table and Credits table. This is because the primary key, account number, in Segment table would be calculated twice. If we mine the data deeper, there are even 6 phone numbers in these 25 phone numbers belonging to different segments. These special mobile numbers are: 1234573570, 1234580271, 1234584022, 1234584602, 1234586598, 1234586874. The schematic diagram for this relationship is shown in **Figure 1**. The mobile one in **Figure 1(a)** illustrates one of the six mobile numbers. This mobile number corresponds to two accounts numbers in Revenue_overage table, one credit in Credits table and two segments in Segment table. It means we couldn't determine whether this credit belongs to segment one or segment two. The relationship of the other 19 mobile numbers is given in **Figure 1 (b).** Because these mobile numbers or credits correspond to only one segment, there is no belonging issue for these numbers in question one.



(a)



(b)

Figure 1. Schematic diagram for the relationship of account number, mobile number, credits and segment: (a) the six special mobile numbers; (b) the other 19 mobile numbers corresponding to two account numbers

If we check the mobile numbers in the Credits table and Revenue_overage table, it can be found that there are 376 numbers only in the Credits table not in the Revenue_overage table, and there are 12 numbers only in the Revenue_overage table not in the Credits table. For the former situation, it means we can't determine which segment these mobile numbers or credits belong to. This is because these mobile numbers don't correspond to any account numbers in Revenue_overage table. As for the latter situation, it means some customers didn't contest their bills. The schematic diagram for this relationship is illustrated in **Figure 2**. There are 14,783 phone numbers in both tables.

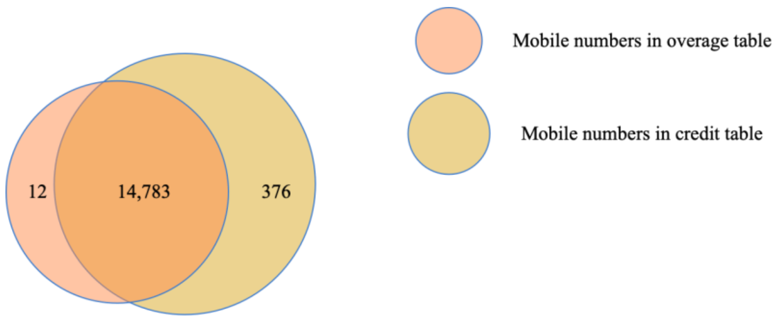


Figure 2. Schematic diagram for the mobile numbers in Revenue_overage table and Credits table

After these analyses, we know that the segment belonging of the 382 mobile numbers (376+6) is unknown. There are 25 phone numbers corresponding to two accounts. We need to deal with these issues in the code (leakage_analysis.py module). There are two ways to deal with these data. One way is to delete or fix these data if it is determined from business domain that these data are incorrect. For example, if we got information from the business department that one mobile number could only correspond to one account number, we need to delete these data. The other way is to put the credits of these 382 mobile numbers in a separated segment (unknown segment). Because there is limited business information for now, the second strategy would be used.

As for the deletion of outlier points, some data might be deleted if more business information is provided. For example, it is noticed that one customer's overage was around 300 dollars, but the credit return was over 20,000 dollars. I am not sure whether it is a data outlier, or it is indeed possible in real life. It is assumed that there are no outlier data in this project considering the given business information.

Another thing is noticed in the initial data analysis. It is found that there are six months data (January to June) for the Revenue_overage table and seven months data (January to July) for the Credits table. Although many data analyses, in my experience, have the same time span for the features, it is assumed that this situation is possible in this project for now. This is because some customers may have their bill overage contest as soon as the bills come out, but many others may begin their contest after some time. Therefore, this is an acceptable assumption. However, if more information is obtained, this assumption could be changed.

For this project, another thing we need to clarify is the policy function (see the policy.py module). In the code, the maximum credit return, based on the TELUS credit policy, is calculated. If this maximum policy credit return is smaller than the real credit return, it would lead to revenue leakage. The maximum credit return, based on policy, can be given as:

$$\text{Max Policy Credit} = \text{policy(overage)} = \begin{cases} 0.5 \times overage \ (\text{overage}<\$1000) \\ 0.5 \times overage \ (\$1000 \le \text{overage} \le \$5000) \\ 0.8 \times overage \ (\text{overage}>\$5000) \end{cases} \quad (1)$$

If the real credit return in Credits table equals to the calculated max credit, the leakage is 0. The overage information can be obtained from the Revenue_overage table. The relationship between max credit and leakage is written as:

$$\text{Revenue Leakage} = \text{Real Credit - Max Policy Credit} \quad (2)$$

3. Discussion and Results

3.1. Hypothesis and Segment Leakage

The hypothesis is correct. The calculated max policy credit and the real credit in each segment are presented in **Figure 3**. The red column is the real credit approved by representatives, and the green column is the max allowed policy credit return. The highest real credit return, around 1.85 million dollars, takes place in the small business solution segment. If we don’t want to have the leakage for this segment, the max allowed credit return, based on the policy, is around 1.32 million dollars. It can be seen that the leakage takes place in every segment except for the TQ – ENT SOLN segment.

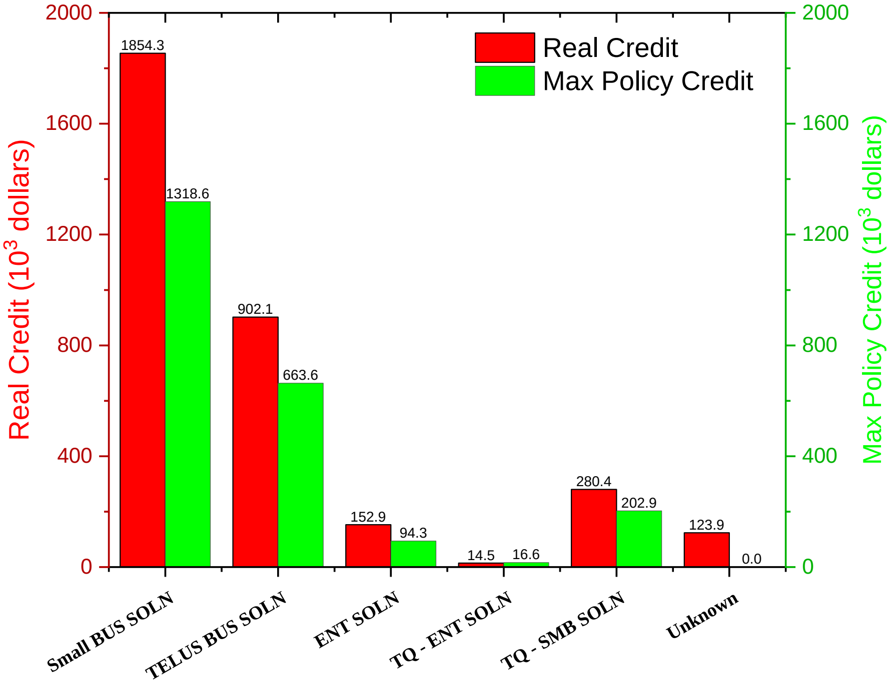


Figure 3. The calculated max policy credit and real credit return for each segment (the unit of the y axis is 10^3 dollars). If the y value is 1200, it means 1200×10^3 dollars = 1.2 million dollars.

Table 1 shows the revenue leakage for each segment. There is no leakage for the TQ-ENT SOLN segment because the real credit return is smaller than the max policy credit return. The leakage for TQ-ENT SOLN is around -2127 dollars, and put it as 0. The total revenue leakage reaches as much as one million.

Table 1. Revenue Leakage generated from each segment

Segment	Small BUS SOLN	TELUS BUS SOLN	ENT SOLN	TQ - ENT SOLN	TQ - SMB SOLN	Unknown	Total
Leakage (dollars)	535,649.64	238,581.60	58,604.36	0	77,564.44	123,913.55	1,034,313.58

The **Figure 4** depicts the ratio of each segment leakage. It can be seen that over 50% revenue leakage comes from the small business solution segment. The unknown segment comes from 382 mobile numbers mentioned in the last section. These numbers didn’t correspond to any account numbers or segments. However, it still results in revenue leakage.

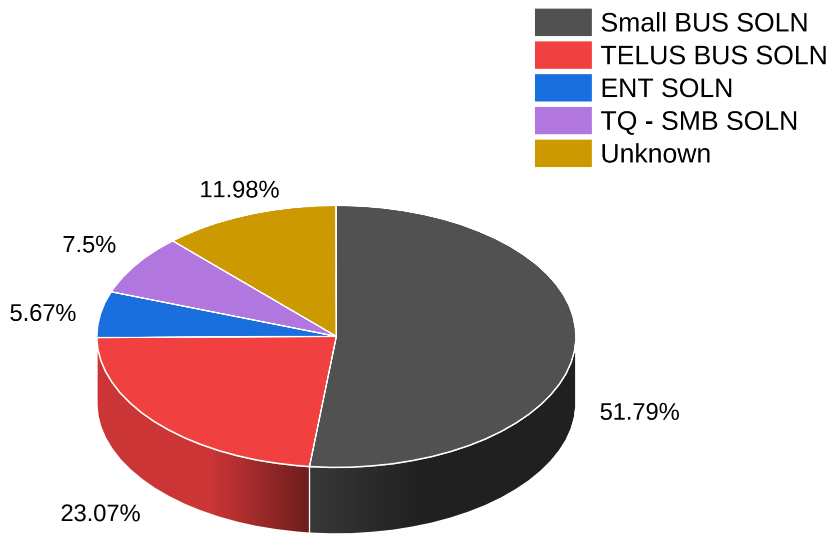


Figure 4. The ratio of revenue leakage for each segment.

3.2. Trend for Overage and/or Credits

There is a trend for the overage and credits. The **Figure 5** shows the relationship among overage, max policy credit return and real credit return each month. It can be seen that the trend of overage is almost the same as max policy credit (same curve shape). The only difference is that the overage is higher than max policy credit in the y-axis. The max credit return is equal to policy(overage) in Eq.(1), and the policy function is a linear function. After calculation, it is verified that the max credit is very close to 0.5×overage when the time scale is month. This is because most customers' overage is less than $5000 dollars. Because this report focuses on leakage analysis (leakage = real credit – max policy credit), the max credit return, instead of overage, would be used and compared with real credit return (Eq. (2)) in other figures. The overage could be further directly derived from the max credit return. It can be seen the highest real credit return takes place one month later than the highest overage and max policy credit. This delay is understandable. The customers' monthly bills usually come out at the end of this month or the early time of next month. The bill shock overage contest could only happen when the bills come out. This leads to the one-month delay of the real credit return compared to the overage. Meanwhile, the trending of overage can affect the trending of real credit return. It means higher (or lower) monthly overage could lead to higher (or lower) monthly real credit. From **Figure 5**, it can also be found there is a revenue leakage every month except for the second month.

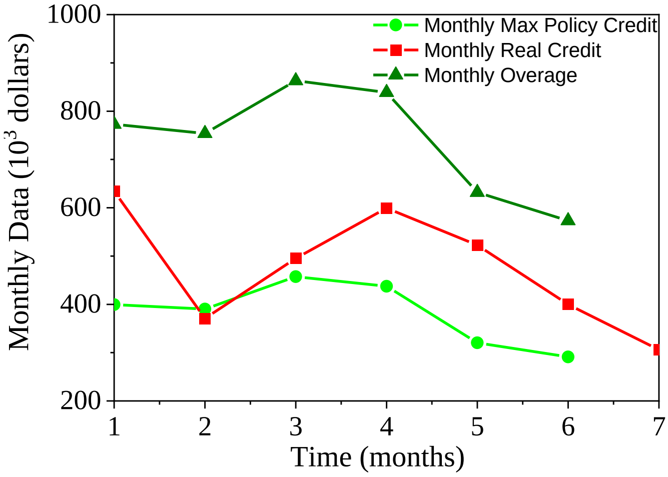
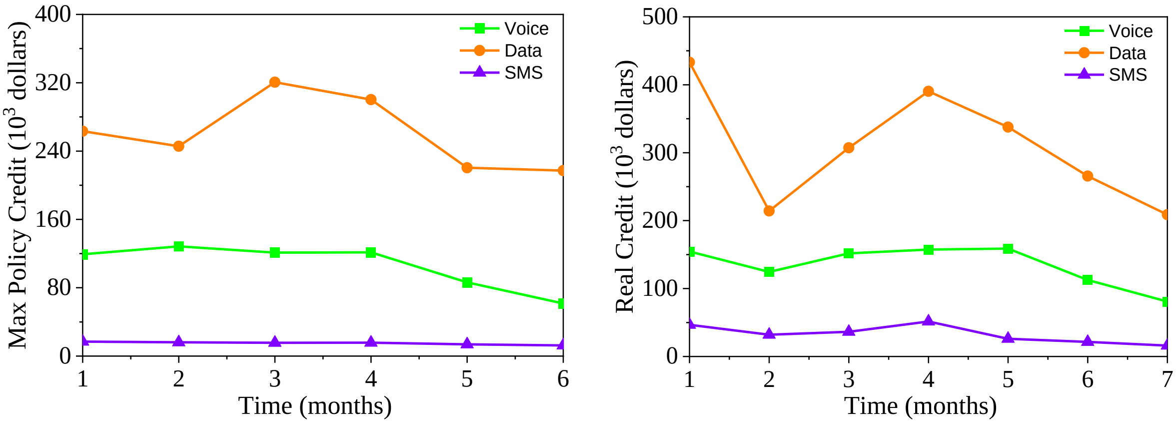
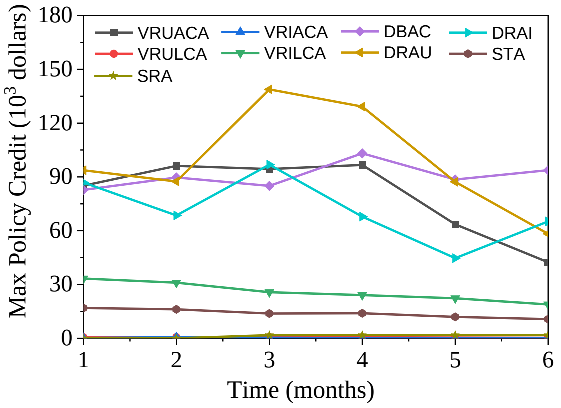


Figure 5. The relationship among overage, max policy credit and real credit every month.

The results for other trending analyses are presented in **Figure 6**. The categories, Voice, Data and SMS, can be checked in **Table 2**. It can be found, in **Figure 6 (a)** and **(b)**, that the monthly max policy credit (or overage) and real credits for SMS are almost constant. By comparing **Figure 6 (a)** - **(b)** and **Figure 5**, it seems that the Data's trending determines the curve's shape in **Figure 5**. Meanwhile, by comparing **Figure 6 (a)** and **(b)**, it can also be found that the one-month delay of real credit. The highest max policy credit (or overage) point for Data is March, and the highest real credit point for Data is April. The max policy credit (or overage) begins to decline from April, and the real credit starts to decrease from May. The SMS curves for both max policy credit (or overage) and real credit are almost horizontal. As for **Figure 6 (c)**, **(d)** and **(e)**, it's hard to find very clear trending patterns.

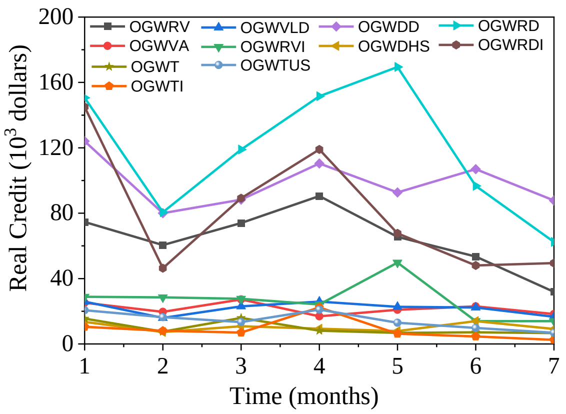


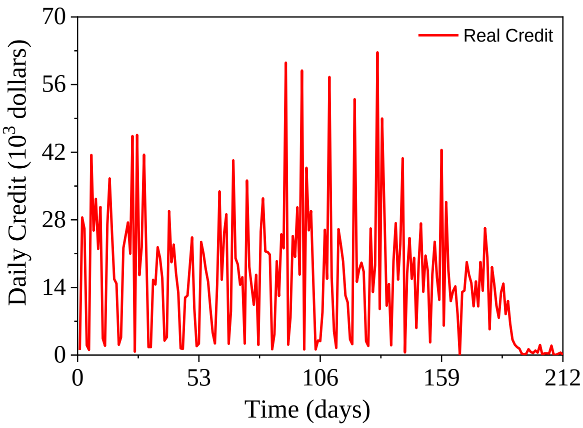
(a) **(b)**



VRUACA: VOICE_ROAM_USA_AIR_CHRG_AMT
VRULCA: VOICE_ROAM_USA_LD_CHRG_AMT
VRIACA: VOICE_ROAM_INTL_AIR_CHRG_AMT
VRILCA: VOICE_ROAM_INTL_LD_CHRG_AMT
DBAC: DATA_BILLED_AMT_CDA
DRAI: DATA_ROAM_AMT_INTL
DRAU: DATA_ROAM_AMT_USA
STA: SMS_TOTAL_AMOUNT
SRA: SMS_ROAM_AMOUNT

(c)





(d) **(e)**

Figure 6. The change of credit and max policy credit with time: (a) max policy credit return for SMS, Data and Voice; (b) real credit return for SMS, Data and Voice; (c) max policy credit return every month for each service; (d) real credit return every month for each service; (e) real credit return every day. The abbreviated legend has been explained above and in the Credits table.

3.3. Highest Source Bill Overage and its Correlation to Credits

The highest source of bill overage is: DATA_ROAM_AMT_USA (DRAU) as shown in **Figure 7**. It is around 1.11 million dollars, and the max policy return is around 0.595 million dollars. The DATA_BILLED_AMT_CDA (DBAC) service is very close. Meanwhile, **Figure 7** also shows the relationship between max policy credit and overage clearly. The overage is almost two times larger than the max policy credit. Although the policy function (Eq. (1)) shows that the coefficient is 0.8 when the overage is larger than \$5000, most customers' overage is less than \$5000. This is the reason that the coefficient is very close to 0.5 considering the used data. Compared with the overage data, the max policy credit should be visualized with the real credit data so that more information could be obtained. Meanwhile, the overage data could be directly obtained through max policy credit data.

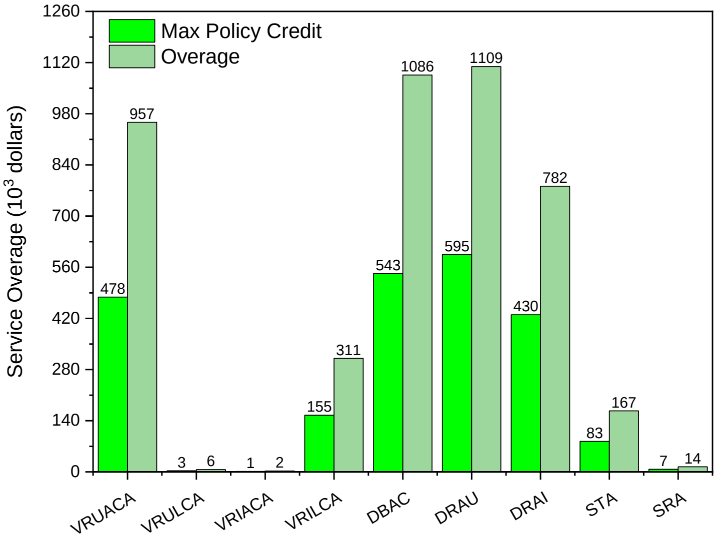


Figure 7. The source of bill overage and max policy credit. The abbreviated labels have been given in Figure 6 (c)

The source of the real credit is shown in **Figure 8**. The highest source of the real credit is the Overage Bill Shock US Roam Data. The source is the same as the overage data or the max policy credit data in **Figure 7**.

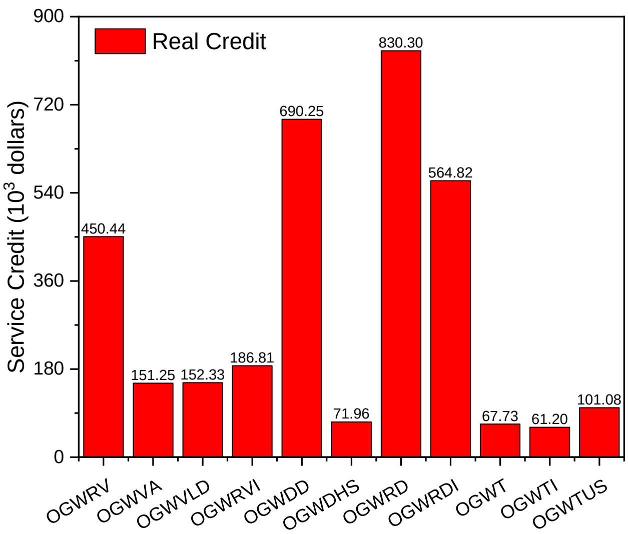


Figure 8. The source of real credits for each service. The abbreviated labels have been given in Credits table

If we further put the source of max policy credit return and real credit return in the same figure, more information about the correlation could be obtained. Before that, one thing needs to be clarified first. It is the category of the services in Revenue_overage table and Credits table. **Table 2** shows the corresponding categories. It shows that the services can be grouped into three categories: Voice, Data and SMS. The Voice includes the Voice USA and Voice International. The Data includes Data Canada, Data USA and Data International. The SMS includes SMS Total (Canada to Canada and Canada to International) and SMS Roaming. Then these categories can be further divided into the services presented in Revenue_overage table and Credits table. In this way, the max policy credit return and the real credit for different services could be plotted in the same figure. The correlation between max policy credit (or overage) and real credit could further be obtained.

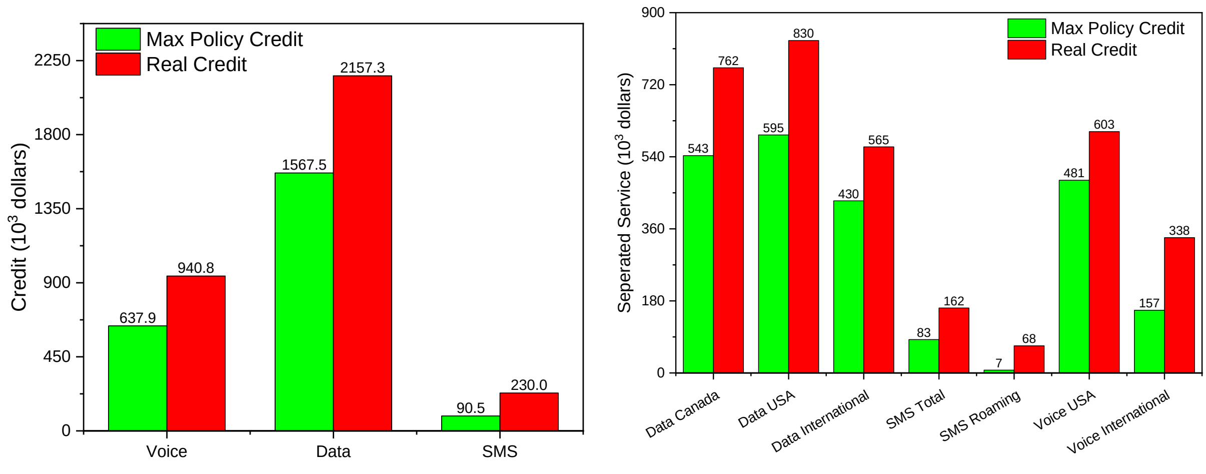
Table 2. service category

Voice	Voice USA	VRUACA	OGWRV
		VRULCA	OGWVLD
	Voice International	VRIACA	OGWVA
		VRILCA	OGWRVI
Data	Data Canada	DBAC	OGWDD
			OGWDHS
	Data International	DRAI	OGWRDI
	Data USA	DRAU	OGWRD
SMS	SMS Total	STA	OGWT
	SMS Roaming	SRA	OGWTI
			OGWTUS

The first level category is depicted in **Figure 9 (a)** and the second level category is shown in **Figure 9 (b)**. It can be found that larger policy credit leads to larger real credit and larger leakage (real credit - max policy credit). Meanwhile it is known that the overage is almost twice as much as the max policy credit (**Figure 7**). It means the overage is almost perfectly correlated to the max policy credit. The Pearson correlation coefficient is used to describe the correlation between max policy credit and real credit. The Pearson correlation can be expressed as:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{3}$$

where r is the correlation coefficient; x_i is the values of the x-variable in a sample, in this case, it is max policy credit of different service. y_i is the value of the y-variable in a sample, in this case, it is the real credit return of different service. $\bar{x}$ is the mean of the values of the x-variable and $\bar{y}$ is the mean of the values of the y-variable. The calculated correlation coefficient of first category (**Figure 9 (a)**) is 1. It means there is perfect positive correlation between the max policy credit and real credit. The calculated correlation coefficient of second category (**Figure 9 (b)**) is 0.9884. It is also very close to the perfect positive correlation. Because the correlation between overage and max policy credit is almost perfect positive. Therefore, the correlation between overage and real credit is also almost perfect positive. It means that greater (or smaller) overage could lead to greater (or smaller) real credit.



(a) **(b)**

Figure 9. credit vs. service for max policy credit and real credit: (a) category level one; (b) category level two

The correlation between overage and real credit has been obtained. The objective is to find the relationship between overage (or max policy credit) and revenue leakage.

3.4. Revenue Assurance for Overage and Credits

As for **Figure 9 (a)** in last section, it can directly derive the leakage (real credit – max policy credit) for the first level service: Voice, Data and SMS. The relationship between max policy credit and identified revenue leakage can be further shown in **Figure 10** (Revenue Leakage = 0.305 × Max Policy Credit + 110.44 × 10^3). This is a very interesting result because the revenue leakage is linear to max policy credit using the data of first level service. Meanwhile, we had the conclusion that the overage was proportional to max policy credit (**Figure 7**). The relationship between revenue leakage and the overage can be further derived as:

$$\text{Revenue Leakage} = 0.15 \times \text{Overage} + 110 \times 10^3 \quad (4)$$

Based on the relationship in Eq. (4), there are three major recommendations to give. If greater overage leads to greater revenue leakage, it means that we need to control the overage. There are several methods to control the overage. One is to send text or email to customers to remind the overage. The other way is to reduce the data speed or stop the corresponding service if the overage takes place. The customers can order an extra add-on for Data, Voice or SMS to continue using the service. Another recommendation is that we can offer a tiered pricing model. It means we could offer the SMS, Voice and Data services with more price points. We could send text, email or give phone call to the customers who had overage several time. We could suggest the customers to change their service plans. Last but not the least, it can be found in Eq. (4) that the revenue leakage would still exist (110,000 dollars) even if the overage is zero. That means some TELUS representatives give customers credits even if they have no overage. To deal with this issue, it requires the automation of the system. The system can decline the representatives' credit request automatically if the real credit is greater than the max policy credit or it requires the approval of the representatives' supervisor for more credits.

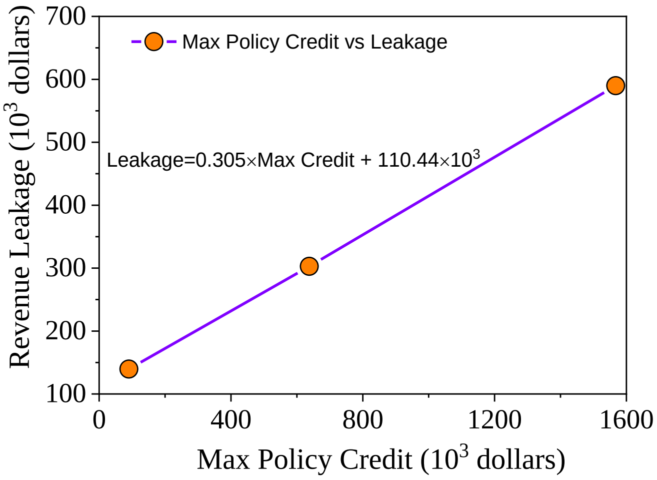


Figure 10. The relationship between max policy credit and revenue leakage.

4. Conclusion

The conclusion presents the direct answers to those four questions in this project. The detailed results and discussion have been given in the previous section.

As for question one, the hypothesis is correct. The size of EBITDA leakage for each segment was given in **Table 1**. All the segments have revenue leakage except the TQ - ENT SOLN.

As for question two, there is a trend in the overage and credits provided. There is one-month delay of credit curve compared to the overage curve in the trending (eg. the highest point in **Figure 5**). This delay is understandable. The customers' monthly bills usually come out at the end of this month or the early time of next month. The bill shock overage contest could only happen when the bills come out. This leads to the one-month delay of the real credit return compared to the overage. Meanwhile, the monthly overage and real credits for SMS are almost constant. The data's trending determines the combined service's curve shape (see **Figure 5** and **Figure 6 (a)**, **(b)**). The trending of overage can affect the trending of real credit return. It means higher (or lower) monthly overage could lead to higher (or lower) monthly real credit.

As for the question 3, the highest source of bill coverage is DATA_ROAM_AMT_USA (DRAU) as shown in **Figure 7**. It is around 1.11 million dollars. The highest source of real credits is Overage Bill Shock US Roam Data. It is the same as the overage data or the max policy credit data. There is almost perfect positive correlation between bill overage and bill shock credits. The Pearson correlation coefficient is almost one. It means highest source of bill coverage corresponds to highest source of bill shock credits.

As for the question 4, there are three major recommendations to give. Firstly, if greater overage leads to greater revenue leakage, it means that we need to control the overage. There are several methods to control the overage. One is to send text or email to customers to remind the overage. The other way is to reduce the data speed or stop the corresponding service if the overage takes place. The customers can order an add-on for Data, Voice or SMS to continue using the service. Another recommendation is that we could offer a tiered pricing model. It means we could offer the SMS, Voice and Data services with more price points. We can send text ot give phone call to the customers who had overage several time. We could suggest the customers to change a service plan. Last but not the least, it can be found in Eq. (4) that the revenue leakage still exists even if the overage is zero. That means some TELUS representatives give customers credits even if they have no overage. To deal with this issue, it requires the automation of the system. The system can decline the representatives' credit request automatically if the real credit is way greater than the max policy credit or it requires the approval of the representatives' supervisor for more credits.

Reference

[1] Abbasi, Wisam, and Adel Taweel. "Provenance-Based Root Cause Analysis for Revenue Leakage Detection: A Telecommunication Case Study." International Provenance and Annotation Workshop. Springer, Cham, 2018.

[2] Curbera, Francisco, et al. "Business provenance–a technology to increase traceability of end-to-end operations." OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, Berlin, Heidelberg, 2008.